THE UNIVERSITY
of EDINBURGH

Detection of signatures of selection in commercial chicken lines

John Joseph Stainton

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Edinburgh

2015

TABLE OF CONTENTS

DECLARATION

I declare that this thesis is my own composition and that the work described within is my own. Specific contributions of others are acknowledged. This work has not been submitted for any other degree or professional qualification.

John Stainton

January 2015

ACKNOWLEDGEMENTS

Primarily, I would like to thank my supervisors, Dr Pam Wiener, Professor Chris Haley, Professor Brian Charlesworth, Dr Andreas Kranis and Dr Kellie Watson for their support and input over the last four years. I would also like to acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC), Aviagen and The Knowledge Transfer Network (KTN UK) for the PhD studentship.

The experimental chapters would not have been possible without the data provided by Aviagen and Hy-Line. I would like to again thank Dr Andreas Kranis for his explanation of the various datasets and the additional information provided. I am indebted to Dr Andy Law and Dr Kai Zeng for their help and support with the programming aspect of the project. I would like to thank Professor Dave Burt for providing invaluable information on the chicken genome and Professor Mike Clinton for his advice as the head of my thesis committee.

I would like to thank my family as well as my friends and colleagues at Roslin for their support and advice over the course of my project. My time at Roslin would not have been as interesting or enjoyable without the highly entertaining lunchtime discussions. I would particularly like to thank Debby, Yan, Eva, Carys, Lucy, Sam, Graham, Brendan, Kate, Fiona, Andrew, Sophie, Zeenath and Smaragda. Finally, I would like to thank Ashley for her support over the four years of this project.

LIST OF PUBLICATIONS

Stainton, J. J., Haley, C. S., Charlesworth, B., Kranis, A., Watson, K. and Wiener, P. (2014) 'Detecting signatures of selection in nine distinct lines of broiler chickens', *Animal Genetics*, **46**(1), 37–49 (Based on Chapter 2)

ABSTRACT

Within the last 100 years, commercial chickens have been split into two main groups. Broiler chickens are produced for meat production while layers are produced for egg production. This has caused large phenotypic changes and the genomic signatures of selection may be detectable using statistical techniques. Genomic regions identified by these techniques may include genes associated with production traits, and is therefore of interest to animal breeders. This thesis investigates signatures of selection in a number of commercial chicken lines using several statistical techniques based on population differentiation and levels of genetic diversity.

First, signatures of selection were investigated using population differentiation in nine lines of broiler chickens. Weir and Cockerham's pairwise F_{ST} was calculated for genome-wide markers between the broiler lines and averaged into overlapping sliding windows to remove stochastic effects. A chromosome bound, circular permutation method was used to generate a null distribution and determine the significance of each window. A total of 51 putative selection signatures were found shared between lines and 87 putative selection signatures were found to be unique to one line. The majority of these regions contain peak positions for broiler QTL found in previous studies and eight regions were significantly enriched for broiler QTL. One region located on chromosome 27 contained 39 broiler QTL and 114 genes, several of which were functional candidates for association with broiler traits.

Secondly, areas of low diversity were investigated in three different SNP datasets. All three datasets were taken from the same broiler line at different time points and consisted of different SNP densities, including 12k, 42k and 600k. A number of zero

diversity regions were found in each dataset and several were shared between the datasets. The 600k dataset was also analysed using a regression test, which investigates the patterns of diversity as the distance from the selected site increases. This method searches for signatures of selections by fitting a regression to the diversity data to test the fit of the data to the theoretical model. A total of 15 regions were found displaying significant asymptotic regression and diversity values less than 0.005. One of these regions located on chromosome 1 was also found as a fixed region in the 12k and 42k datasets and contained the gene *IGF1*, which encodes an important protein for growth.

Finally, signatures of selection were investigated between broiler and layer datasets by investigating population differentiation and diversity based analysis. Weir and Cockerham's pairwise F_{ST} was calculated between the two lines and outliers extracted. A total of 32 regions were found displaying high differentiation. Seven regions of low diversity in the layer dataset were also investigated. Several broiler and layer QTL had been previously identified in these regions. Two genes related to hedgehog proteins were identified within selected regions, which are known to be involved in embryogenesis. Finally seven regions were found to be highly differentiated between the broiler and layer lines, and the nine broiler lines in the first chapter. This may indicate selection which occurred during breed separation.

Signatures of selection were identified in four broiler and layer datasets using several statistical techniques. A number of regions were identified in multiple datasets by a number of techniques and are therefore good candidate regions for selection. Other statistical techniques could be used in future studies to further confirm these regions and identify causative genes and variants.

LAY ABSTRACT

Chickens are an important source of food as they produce both meat and eggs. Over time, lines of chickens specialised to produce meat more efficiently have been created using artificial selection. These populations are known as broiler chickens. There are also populations specialised to produce eggs, which are known as layer chickens. This selection has caused large changes to the physical characteristics of the chickens and there may be signs of this selection present on the genome of these animals. Regions on the genome displaying signatures of selection may contain genes associated with meat or egg production.

In order to find these genes related to production, signatures of selection were investigated in two main ways across several chicken populations. The first method is population differentiation, which is used to find significant differences in the genome between two or more populations. This was used to find signatures of selection by comparing nine different populations of broiler chickens and by comparing a broiler population to a layer population. The second method looks for areas in the genome which display low diversity in the population. This was used in several populations of broiler chickens, as well as one layer population. A number of signatures of selection were found for each population, some of which contain genes known to be associated with production traits. Other regions do not contain obvious candidate genes and need to be investigated further using other statistical techniques.

CHAPTER ONE

General Introduction

1.1 Introduction

The domestication of animals is an important part of human history and began approximately 11,000 years ago (Zeder, 2008). This process allowed the creation of animal breeds with useful traits, which has had several advantages for humans including an increased rate of food production. This is essential, in view of the constantly expanding human population. Examples of some recent results of selective breeding of domesticated animals include increased meat and milk in cattle (Grobet et al., 1997), increased meat production in salmon (Gjederm, 1979) and the creation of broiler and layer lines in chickens (Muir et al., 2008).

Recent advances in genomic technology, such as whole genome sequencing and high density SNP panels, allow the domestication process to be investigated in depth. Several methods suitable for the investigation of domestication traits are available. QTL mapping may be used to identify the loci responsible for a trait of interest, but this approach requires extensive phenotypic information (Haley, 1995). A complementary approach investigates regions in the genome which are displaying signatures of selection. This approach may be particularly appropriate in domesticated animals, due to the strong artificial selection in the domestication process.

1.2 Selection

Traits which improve fitness, either by increasing an organism's chance of survival or reproductive success, are more likely to be passed on to the next generation and therefore increase in frequency in the population. This process is known as natural selection. Selection criteria can be imposed on an animal species by humans to produce animals with favourable traits. This is known as artificial selection and it has similar

properties at a genetic level to natural selection. There are several different types of selection, which are defined below.

1.2.1 Positive Selection

Selection often acts in a directional fashion, favouring a steady increase in a mean phenotype. Under such "positive selection", a particular allele is favoured and spreads through the population. Positive selection is understood to be the main mechanism of adaptation and is a key element in animal domestication and artificial selection. This mode of selection can leave conspicuous signals on the genome, known as selective sweeps. Neutral sites close to the positively selected site are dragged along with the favoured allele , resulting in an elimination of variation at neighbouring sites in the genome (Maynard Smith and Haigh, 1974). Selective sweeps can either be hard sweeps, where the selected allele reaches fixation, partial sweeps, where the allele frequency increase rapidly but does not reached fixation, or soft sweeps, where standing genetic variation becomes advantageous and is positively selected, rapidly increasing in frequency (Przeworski et al., 2005). Soft sweeps may also refer to a situation where several new mutations at a single locus are all equally advantageous and increase in frequency simultaneously, preventing one specific allele from becoming fixed.

As well as the reduction in variation around the selected site, inflated linkage disequilibrium can occur and the frequency spectrum shifts from a neutral model to a surplus of rare mutations in the population (Tajima, 1989). Additionally, if two or more populations are subjected to different selection pressures, they may become differentiated with respect to allele frequencies at both the selected locus and loci linked

to it (Weir and Cockerham, 1984). These effects of selection may be detected by several statistical techniques, which are described below.

1.2.2 Negative selection

With negative (or purifying) selection, a mutation is at a selective disadvantage and is eventually removed from the population. Random mutations are often deleterious, so that many novel alleles are immediately subjected to negative selection and are quickly removed from the population. This can result in regions of reduced variability in the genome where strong background selection is present.

1.2.3 Balancing selection

Balancing selection maintains multiple alleles at a locus in the gene pool. This can occur in cases of heterozygote advantage, where a heterozygous individual at a particular locus has greater fitness than a homozygous individual. The most famous example of this is sickle cell anaemia in humans, which also confers malaria resistance. A homozygous individual will experience the full effect of the disease, but the heterozygous individual will gain malaria resistance but not experience many ill effects. Therefore, individuals with heterozygous alleles are very common in areas with persistent malaria outbreaks (Allison, 1954). Balancing selection can also occur due to frequency-dependent selection where the fitness of a phenotype is based on its frequency relative to other phenotypes in the population.

In most cases, only positive selection is relevant in the context of artificial selection and animal domestication. However, balancing selection can be important if an allele that is favoured by artificial selection is deleterious when homozygous. This can be seen in Dexter cattle which suffer from chondrodysplasia in homozygous individuals (Harper et

al., 1998). This condition leads to dwarfism and the calves are often aborted before the seventh month of gestation.

1.3 Detecting Signatures of Positive Selection

There are a number of tests that detect patterns generated by selective sweeps. These are detailed below.

1.3.1 Linkage Disequilibrium

Linkage Disequilibrium can be generated by selective sweeps created due to the hitchhiking effect (McVean, 2007). Selection acts on individuals with advantageous mutations which gain a reproductive advantage or increased survival. Selective sweeps include neutral variants which are linked to the beneficial mutation and so are selected alongside them. This creates long haplotypes in the genome which are broken up by recombination over time. Therefore, the presence of a long and frequent haplotype suggests that there is recent or ongoing selection in the population.

Linkage disequilibrium is commonly measured by the D' statistic, which is the absolute value of D, relative to its maximal value conditioned on the allele frequencies (Lewontin, 1964) and the r^2 measure, which is the squared correlation coefficient of the frequencies at two loci (Hill and Robertson, 1968). The r^2 measure was used to investigate LD in nine lines of broiler chickens (Andreescu et al., 2007). Widespread LD patterns were found, but over shorter distances than found in other livestock species. The LD patterns are consistent across closely related lines. However, this measure can be biased if the genotyped individuals are closely related (Mangin et al., 2012).

A number of LD based tests are designed to detect long haplotypes relative to their frequency within a population. These regions must have become common rapidly, or

recombination would have begun to break down the LD and the haplotype would have become shorter. There are several tests based around the extended haplotype homozygosity (EHH) statistic (Sabeti et al., 2002). EHH is defined from a core region, such as a putatively selected allele, to a specific distance in both directions. The probability that two randomly chosen chromosomes carrying the core region are identical by descent is calculated. EHH decreases the further we travel from the core region. This reflects the action of recombination breaking down the haplotype in the population over time (Sabeti et al., 2002).

The long-range haplotype (LRH) test can detect recent positive selection by comparing a haplotype's frequency to its EHH at several distances from the core region (Sabeti et al., 2002). Long haplotypes with a high frequency in the population suggests that recombination has not had long enough to break down the haplotype, and therefore the haplotype rose to high prevalence quickly. This may be a sign of positive selection.

EHH has been investigated in a number of species. In chickens, it has been used to investigate selected regions in two lines divergently selected for abdominal fat content (Zhang et al., 2012b). Ten genes were implicated for abdominal fat content, seven of which were in previously reported QTL regions for fatness. In white leghorn chickens, a total of 186 genes were found in regions demonstrating linkage disequilibrium (Li et al., 2012). Many of these genes were associated with egg production and growth traits.

A variant of the EHH test is the integrated haplotype score (iHS) test. This test looks at the area under the curve of EHH values as the distance from the core region gets larger (Voight et al., 2006). Positive selection is detected by calculating the area under the curve. Alleles experiencing positive selection are likely to display extreme EHH values a

short distance from the core region and moderate EHH for longer distances. This technique is most useful for detecting incomplete sweeps.

The iHS test has been used in a number of previous studies. One study investigated the convergent adaptation of the lactase persistence gene in European and African human populations (Tishkoff et al., 2007). The C-14010 allele was associated with lactase persistence in African populations and was found to have extreme iHS score relative to simulated data under several models. In cattle, the both the iHS and F_{ST} tests (described in the population differentiation section below) were used together to investigate positive selection in dairy and beef breeds (Qanbari et al., 2011). One selection signal displayed both high differentiation and extreme iHS values near the *Sialic acid binding Ig-like lectin 5* gene, which has previously been associated with longevity and calving ease in Holstein cattle. Another study focused on chromosome 6, and found a cluster of iHS values around the *ABCG2* gene (Hayes et al., 2008). This gene contains a mutation affecting protein percentage, as well as other production traits.

The ln(Rsb) test is another method based on EHH and is designed to complement iHS (Tang et al., 2007). The iHS test compares EHH between alternative alleles at a SNP. Therefore, it has little power when one allele is at a high frequency. To overcome this problem, the ln(Rsb) test compares EHH at the same SNP in two different populations. This test has high power for detecting selection where the selected allele has reached fixation. This test was applied to SNP data from human populations and found a number of genes previously found to be under selection, including *CASP12*, a gene involved in disease resistance. However, this approach failed to detect certain genes known to be under selection, as these regions were subject to partial sweeps.

Another variant of this test is the cross population extended haplotype homozygosity (XP-EHH) (Sabeti et al., 2007). The previous two tests in this category detect positive selection by finding long haplotypes with high frequency in the population. However, the XP-EHH is specialised in detecting sweeps in which the selected allele is fixed (or close to fixation) in one population, but is polymorphic in the overall species population. Haplotype lengths are compared between populations to control for local differences in recombination rates.

The XP-EHH and the iHS tests were used to investigate adaptation to high altitude environments in Tibetan human populations (Simonson et al., 2010). The XP-EHH test identified six candidate genes, while the iHS identified five. Both tests identified *EGLN1* as a candidate, which is associated with haemoglobin levels. A number of statistical tests, including the XP-EHH, were also applied to detect signatures of selection on the X chromosome in three breeds of pigs (Ma et al., 2014). A total of 29 regions were found by the test across the three lines, including three genes associated with haematological traits and the *ACSL4* gene, which is associated with meat quality.

Finally, the LD decay (LDD) test was created to avoid the process of determining haplotypes (Wang et al., 2006). This is achieved by focusing on SNP loci which are homozygous in the population. The fraction of inferred recombinant chromosomes (FRC) at adjacent SNPs can be directly computed. Differences in local recombination rates will theoretically affect this test less, as the method investigates the LD differences between both alleles rather than the extent of LD around the alleles.

1.3.2 Local Reductions in Genetic Variation

Positive selection can cause an advantageous allele to spread throughout the population and become fixed. This causes a reduction of diversity in the population at the selected site. The hitchhiking effect reduces genetic variation at neutral sites linked to those under positive selection, and causing an area of low variation around the selected gene (Maynard Smith and Haigh, 1974). Variation will increase over time due to the effect of new mutations and recombination although if the allele and linked sites have gone to complete fixation, recombination has no effect. This effect should be detectable by calculating the diversity of all areas of the genome in the population and investigating regions with particularly low variation. A study in human populations used this reduction in diversity as well as population differentiation in order to identify 180 regions with strong evidence for selection (Oleksyk et al., 2008).

A variation of this technique was used in eight populations of domestic chickens and one population of red jungle fowl where a heterozygosity (H_P) statistic was adapted to be used in pooled data. In order to distinguish between true sweeps from fixation due to genetic drift, the sequence data was pooled into three groups based on a common selection purpose. These groups included domestic lines, two commercial broiler lines and three layer lines. H_P was calculated in 40kb sliding windows. A region of approximately 40kb was found to be nearly fixed at the *TSHR* gene. In addition, screening the broiler population revealed several sweeps related to selection for muscle growth. This includes a region on chromosome 1 which includes *IGF1*, a candidate gene

for growth, and *PMCH*, a gene with roles in appetite and metabolic regulation (Rubin et al., 2010).

The pooled heterozygosity method has been used in several further studies in chickens. The first implemented the H_P statistic in 67 breeds of chickens (Elferink et al., 2012). A total of 26 regions showed strong evidence of selection, 13 of which had been previously described. This study was unable to find the selection signature at the *TSHR* locus as there was only one SNP present in this analysis within the 40kb region found by Rubin et al (2010). This SNP was fixed in nearly all domesticated breeds but the windows that included this SNP did not reach significance as the other SNPs present in the window were segregating at high frequencies. Other additional regions were found, including three regions that were found exclusively in broiler breed groups. Secondly, the H_P statistic was calculated, along with several other statistics, in two lines of chickens divergently selected for abdominal fat content (Zhang et al., 2012a). Several regions were implicated as being affected by selection, including a region on the Z chromosome containing the *PCSK1* gene and a region on chromosome 1 containing the *IGF1* gene. Both genes have previously been associated with fatness in chickens. Thirdly, H_P was calculated in a commercial brown layer line (Qanbari et al., 2012). This was combined with a variant of the sliding window method, referred to as a "creeping window". Eighty two regions with strong evidence of selection were found, containing genes such as *CALB1* on chromosome 2, which is related to egg shell quality, and *SPP1* which is related to egg shell calcification. Finally, this technique has also been used in pooled data from pigs (Rubin et al., 2012). Strong signatures of selection were found at three

loci associated with morphological changes linked to pig domestication, (the elongation of the back and the increased number of vertebrae).

Another method for detecting low diversity regions utilises the diversity pattern relative to the genome position (Wiener and Pong-Wong, 2011). The heterozygosity levels in the population should increase as the distance from the selected locus increases. This approach fits a regression to the heterozygosity data as a function of genomic position, to test the fit of the data to the model predicted by theory (Wiener and Pong-Wong, 2011). This method was compared to the iHS test for detecting selection and was found to have both good accuracy and power. Significant asymptotic regressions were found when this test was applied to the *myostatin* gene region in cattle, responsible for the double muscling phenotype, and the *VKORC1* gene region in rats, responsible for warfarin resistance. An additional advantage of this test is that it can be applied to a number of different types of data. This method was also applied to sheep data to investigate signatures of selection related to dairy production traits, along with pairwise F_{ST} (defined later) and observed heterozygosity calculations (Gutiérrez-Gil et al., 2014). One region on chromosome 6 was found by all three methods and includes the *ABCG2* and *SPP1* genes. This region is orthologous to a region on the bovine genome where several milk production QTL have been found.

However, tests based on a local reduction of diversity have some drawbacks. The local reduction in diversity can persist in the genome for a long time and across a long region in the genome. This can make finding the selected site itself quite difficult. Also, certain demographic events, such as founder events or recent population bottlenecks can also

reduce diversity across the genome (Nei et al., 1975, Templeton, 1980, Johnson et al., 2007)

1.3.3 Changes to the Frequency Spectrum

A selective sweep reduces the variability around a selected site. Eventually, new mutations will begin to appear at low frequencies in the population, as the time taken for a new mutation to increase in frequency under neutral drift is low. This means that there is a surplus of rare alleles in the population (Charlesworth and Charlesworth, 2010). The frequency spectrum will eventually move back to its original state before the sweep, but this may take thousands of generations. The first test developed for detecting this signal was Tajima's D (Tajima, 1989), which compares the mean pair-wise differences between individuals in a population with the total number of polymorphic sites. Negative values of this statistic suggest an excess of rare alleles, which may indicate positive selection. This test can also be used to investigate balancing selection, which has the opposite effect.

A change in the frequency spectrum due to a selective sweep also changes the frequency of ancestral and derived alleles in a region. Derived alleles will be found at low frequencies as they are relatively young compared to the ancestral alleles. Positive selection will cause higher frequencies of derived alleles than would be expected by chance. Several tests take advantage of this process, including Fu and Li's F test (Fu and Li, 1993) and Fay and Wu's H test (Fay and Wu, 2000). However, these tests can be biased due to the ascertainment bias found in most genome wide SNP datasets.

A different method involves the calculation of a composite likelihood ratio (CLR) (Kim and Stephan, 2002, Zhu and Bustamante, 2005, Nielsen et al., 2005, Jensen et al., 2005). These tests also require a null distribution of variation produced via simulations. This type of test has the most power when the selected allele has become recently fixed. In a study in Fleckvieh cattle using a 15 million SNP panel, iHS and a CLR test were applied to investigate signatures of selection (Qanbari et al., 2014). There was a substantial amount of overlap between regions identified by both methods. Candidate genes under selection included *KIT* and *MITF*, which have been shown to affect coat colour in a number of mammals.

Another frequency spectrum based approach uses hidden Markov models (HMMs) (Boitard et al., 2009). This method takes advantage of the distortion of allele frequencies present around a selected site. It uses the full allele frequency spectrum rather than a summary statistic, which increases the test's detection power. This test was applied to 30 unrelated animals from the Blonde d'Aquitaine cattle breed (Boitard and Rocha, 2013). Three candidate regions were found, including one on chromosome 2 containing *growth differentiation factor 8 (GDF8)*, the gene responsible for the double muscling phenotype.

Many studies employ a sliding window analysis instead of using single locus values. This technique removes stochastic variation between loci and improves discovery of signals of selection (Weir et al., 2005). The cross population composite likelihood ratio (XP-CLR) test removes the need for this by modelling the positional patterns of allele frequencies along a chromosome as a function of genetic distance to the selected allele

(Chen et al., 2010). This method is similar to the EHH test which takes advantage of linkage disequilibrium instead of population differentiation. Regions are considered to be under selection if the change in allele frequency occurred too quickly at the selected site to be due to random drift. It is also claimed that this technique is more robust to ascertainment bias, as it is based on allele frequency differentiation across populations. This technique was used to examine genetic variation in human Tibetan populations and investigate high altitude adaptation (Peng et al., 2010). 131 regions were found displaying signatures of selection, containing 179 genes. The top candidate genes are related to hypoxia and include *EPAS1* and *EGLN1*.

1.3.4 Population Differentiation

Two populations of the same species in different isolated environments are likely to experience different local conditions and therefore different selection pressures. If gene flow is restricted between these two environments, the selection signatures in both populations could significantly differ from one another. As well as there being decreased variation within a population, regions where a selective sweep is present should also experience greater variation in allele frequencies between populations. Population differentiation is a method of investigating this variation in allele frequencies between populations. It can be quantified by the statistic, F_{ST} (Wright, 1951). Under selective neutrality, F_{ST} is determined by genetic drift, which will affect all loci across the genome in a similar fashion. When selection is present, F_{ST} will deviate from these expected values. Positive selection may lead to an increase in F_{ST} (Akey et al., 2002, Cruickshank and Hahn, 2014)

One of the first methods to take advantage of population differentiation was the Lewontin-Krakauer test (Lewontin and Krakauer, 1973). This approach was developed at a time where genetic markers were limited. Therefore, the test was focused on specific markers and compared the observed value of inter-locus of the estimate of Wright's F_{ST} with the theoretical variance. The test makes two assumptions: F_{ST} is the same for all loci and the gene frequency of each subpopulation used is a random sample from a given frequency distribution. This test was later criticised for being unreliable, as these assumptions were found not to be justified in all circumstances (Nei and Maruyama, 1975). Despite these criticisms, this test has been expanded upon several times. Many of these extensions attempt to increase the power of the test by accounting for population demographic events or complex population trees and by minimising the false positive rate (Bonhomme et al., 2010).

Later tests involved running simulations to generate a null distribution of F_{ST} values. Outliers could then be identified in the real F_{ST} values and investigated as putative signatures of selection. This technique was used to investigate 100 DNA polymorphisms in five human populations from four continents (Bowcock et al., 1991). They estimated that approximately a third of their dataset contained selected genes. This approach was later expanded to use a coalescent model, which examined the observed distribution of F_{ST} values as a function of heterozygosity. These distributions were compared to expected results from symmetrical island model simulations to identify signatures of selection (Beaumont and Nichols, 1996). Finally, a Bayesian model using Markov Chain

Monte Carlo simulations was implemented, which can identify loci when strong selection is present (Beaumont and Balding, 2004).

These previous methods require a simulation step in order to obtain the expected distribution of F_{ST} under neutrality. However, the simulated distribution of F_{ST} depends strongly on the demographic history of the population, which is rarely reliably known. Because of this problem, the most recent studies involving population differentiation have applied a different approach to investigating significant F_{ST} values. This approach skips the simulation step and instead categorises a large number of SNPs throughout the genome. Loci which are affected by selection should show up as outliers in the tails of the empirical distribution of F_{ST} (Akey et al., 2002, Fullerton et al., 2002). This approach is possible due to the relatively recent availability of high density SNP panels and improved genotyping technology.

This approach was first used in three human populations: African-American, East Asian and European-American using 26,530 SNPs. F_{ST} values were found to cluster together on chromosomes; estimates of F_{ST} for nearby SNPs were correlated. One hundred and seventy four candidate genes were found by mapping 8862 SNPs to gene-associated regions. This includes 17 genes underlying known Mendelian or complex disease traits (Akey et al., 2002).

This approach has been used in a number of other species. Dogs are a particularly important species for this type of study, due to thousands of years of artificial selection. Several candidate genes associated with skin wrinkling in Shar-Pei dogs were suggested

by the application of differentiation tests (Akey et al., 2010). F_{ST} analysis in pigs have revealed regions under selections associated with coat colour, ear morphology and teat number (Wilkinson et al., 2013). In a study of several dairy cattle breeds (including Holstein, Normande and Montbéliarde), genes showing evidence of differentiation included the *Growth Hormone* gene (*GH*) and *Insulin Growth Factor* gene (*IGF1*), which are both important in milk production (Flori et al., 2009).

F_{ST} calculations can be combined with phylogenetic relationship to calculate locus specific branch lengths (LSBL) (Shriver et al., 2004). This is useful when comparing multiple populations, as it reveals which population experienced changes in allele frequency at a particular locus, as well as the amount of change. This technique was used in four human populations, and found significantly higher LSBL values for markers located on the X chromosome. This method was also used to investigate positive selection in western African pygmy populations (Jarvis et al., 2012). Several pygmy exclusive regions were found on chromosome 3, between 45 and 60 Mb. These contain several candidate genes associated with height variation, including *CISH* and *DOCK3*.

1.3.5 Combining Test Statistics

Methods have been developed which combine a number of different tests for multiple signals of selection into one statistic, theoretically increasing the power and resolution of the selection signatures. The first of these combined statistics is the DH test (Zeng et al., 2006), which combines Tajima's D (Tajima, 1989) with Fay and Wu's H test (Fay and Wu, 2000). This test is powerful for detecting positive selection, but frequency spectrum based tests can be very conservative. To address this, the test was expanded upon with

the HEW test, which combines Fay and Wu's H with the Ewens-Watterson (EW) test, and the DHEW test, which combines DH with EW (Zeng et al., 2007). The Ewens-Watterson (EW) test detects positive selection based on haplotype frequency spectrum (Watterson, 1978). Both of these tests were found to be robust to recombination and are useful for detecting recently fixed selective sweeps.

Another of these combined statistics is the composite of multiple signals (CMS) test (Grossman et al., 2010). This method combines signals from long range haplotype based tests (iHS and XP-EHH), population differentiation tests (F_{ST}) and two additional tests developed for use in this statistic. The first of these, ΔDAF, is based on the frequency spectrum and tests for derived alleles found at high frequencies relative to other populations. The second, ΔiHH, is a haplotype based test which measures the absolute length of haplotypes and is more sensitive to selected alleles at lower frequencies.

Coalescent simulations were used to determine each test's ability to distinguish selected variants from neutral markers and to localise the signals of selection spatially. This also accounts for the correlation structure between each individual test. The CMS test was used to localise the selected variant within a candidate region, as it was found that this test has the power to localise the selection signal within 89kb for full sequence data, while individual tests only provided localisation within 1Mb (Grossman et al., 2010). This method was later modified into the genome-wide CMS (CMS_{GW}) method (Grossman et al., 2013), which can be used identify candidate regions within the genome. This modified method identified 86 putative regions under selection in full genome sequencing data from human populations in the 1000 Genomes Project. The

standard CMS test was used to fine map these regions for candidate mutations. Thirty five candidate nonsynonymous mutations were found, as well as a number of candidate regulatory elements.

A separate method, Meta-analysis of Selection Signals (meta-SS), was developed for use in dairy and beef cattle data (Utsunomiya et al., 2013). This method combines signals from a number of tests, including long range haplotype based methods such as iHS and Rsb, Z-transformed heterozygosity, a reduction in local diversity based method and ΔDAF, the frequency spectrum based test developed for use in the CMS test. The P-values from each of these methods were combined using a weighted Z transformed method (Whitlock, 2005). The most significant selection signal found by this method was in the CNIH3 gene in the Brown Swiss dairy breed. This gene regulates the trafficking properties of AMPA receptors in the central nervous system (Shia et al, 2010) which have previous been shown to be involved in luteinizing hormone surge (Brann and Mahesh, 1997) and is essential for reproduction.

The composite selection signals (CSS) test combines three selection tests; F_{ST}, XP-EHH and either ΔDAF or ΔSAF (Randhawa et al., 2014). ΔDAF is the frequency spectrum based test developed for the CMS test (Grossman et al., 2010). ΔSAF is also an allele frequency spectrum based test which takes into account the allele frequency differences between sheep populations, due to the lack of ancestral allele information available in sheep. These tests are combined using the rank distribution of each constituent test. This test was found to have higher accuracy and localisation than using the individual test alone. In one dataset, high CSS results were found in the candidate gene region, while

the XP-EHH test reported a signal 5Mb upstream of the region and the ΔSAF and F_{ST} signals were located 4Mb downstream of the region.

Finally, the decorrelated composite of multiple signals (DCMS) test has been developed very recently (Simianer, 2014). This test avoids using coalescent simulations, like the CSS and meta-SS tests, but also accounts for the correlation structure between the individual statistics, like the CMS test. The power of DCMS was found on average to be 14.5% higher than other individual test statistics, including iHS, EHH and F_{ST}. When compared to other combined methods, the power was found to be 7.2% higher than meta-SS and 26.6% higher than CSS.

1.3.6 Problems with the detection of selection using statistical tests

All of the above statistical tests have their own strengths and weaknesses. However, there are several limitations shared between all these tests. Firstly, SNP data is often used in selection studies. However, all SNP chips suffer from some form of ascertainment bias (Lachance and Tishkoff, 2013). When the SNP arrays are created, rare, low frequency alleles may not be detected and therefore will not be present in the SNP datasets (Gravel et al., 2011). This suggests that the full extent of genetic diversity in the population will not be captured by the SNP array and therefore there will be biases present in the dataset.

Secondly, demographic events can also create signals similar to selection signatures. Population bottlenecks and founder events can cause regions of low diversity in the genome, which will affect the results of all tests. For example, African populations in humans have been found to have increased diversity and low levels of linkage

disequilibrium compared to non-African populations. This reduced diversity in non-African populations is due to a bottleneck associated with dispersal from Africa (Tishkoff and Verrelli, 2003). Population expansions can cause an excess of low frequency alleles compared to the number expected by chance by a neutral model (Biswas and Akey, 2006), similar to the effect detected by some frequency spectrum based tests, such as Tajima's D (Tajima, 1989). It can be difficult to correctly distinguish signals created by these events from true selection.

1.4 Domestic Chickens

The modern domestic chicken (*Gallus gallus domesticus*) is an important food source for current human populations due to the production of both meat and eggs. The ancestor of these birds is a subspecies of the Red Jungle Fowl (Gallus gallus gallus) (Fumihito et al., 1994, Crawford, 1990). There are several other closely related species of junglefowl which may have bred with the red junglefowl in the past, including green junglefowl (*Gallus varius*), grey junglefowl (*Gallus sonneratii*) and ceylon junglefowl (*Gallus lafayetii*) (Crawford, 1990). There is some evidence that several genes, including the gene for yellow skin, were integrated into the domestic chicken via a hybridisation event with the grey junglefowl (*Gallus sonneratii*) (Eriksson et al, 2008). Of these species, the red junglefowl has the largest geographic range and is divided into several subspecies.

The principle purpose of domestication was not for food production. Instead, chickens were domesticated for entertainment and cock fighting (Crawford, 1990). When cock fighting became illegal in the 1800's, chickens were bred for display or religious

purposes and were selected for visual traits such as coloured plumage patterns (Crawford, 1990). During this time, several breeds were developed which became important during the creation of modern broiler and layer breeds. These include the Cornish Game Hen and Plymouth Rock breeds, which are important in broiler breeds because of their particular body conformation and the White Leghorn, which are important in layer breeds because of their high egg production.

The main focus of selection for meat and egg production began in the early 20th century where commercial breeds were strongly selected for production traits (Gholami et al, 2014). The selection process caused significant behavioural changes to facilitate the increase in production traits. For example, the red junglefowl lays approximately 10 to 15 eggs a year in the wild. Current laying hens lay over 300 eggs per year (Moreng and Avens, 1985, Romanov and Weigend, 2001). Leghorns invest more energy in production by roaming less than their wild ancestor (Schütz et al, 2002). Foraging behaviour is reduced and modern breeds show less capacity for learning than the red junglefowl (Väisänen and Jensen, 2004). However, social behaviours, such as pecking order, are maintained in modern chickens (Väisänen et al 2005).

Intensive selection has had a significant effect on nucleotide diversity within modern breeds. One study evaluated genetic diversity by estimating the genetic distances between breeding stock (Siegel et al, 1992). They concluded that the specialised lines for meat and egg production suffer from reduced diversity, but there is considerable amount of genetic diversity in the overall chicken population. Conversely, a study using DNA

fingerprinting found considerable genetic diversity in USA commercial breeder populations (Dunnington et al, 1994).

Chickens may be particularly suitable for the detection of selection signatures as domestic chickens were split into two distinct lines specialised for a specific function approximately seventy years ago. Broilers are used for meat production and layers for egg production. This was accompanied by large phenotypic changes and the signatures of selection are likely to be detectable by statistical techniques. The need for separate lines is due to the negative genetic correlation between selection for growth and fertility (Barbato, 1999). Male broilers originated from Cornish stock, which are from the British Cornish Indian game breed, and have a thick compact body type with a high proportion of breast muscle (Muir et al., 2008). The dams originate from breeds such as White Plymouth Rock and many of the same dual purpose breeds, which are used for brown egg production. Broilers are selected for a large variety of traits. Many of these traits are related to production, such as growth rate, body weight, residual feed consumption, feed conversion ratio and abdominal fat content. However, they are also selected for welfare traits, such as leg health and vaccine response and some reproduction traits. Selection on layers is primarily focused on reproduction traits including egg shell quality, egg shell thickness and incubation behaviour as well as some welfare traits.

The chicken genome was first sequenced in 2004 (International Chicken Genome Sequencing Consortium, 2004). However, large portions of the sequence were missing due to the unique avian karyotype. This karyotype is composed of 38 autosome pairs and two sex chromosomes, Z and W (Smith and Burt, 1998). These chromosomes exhibit a

large variation in size, ranging from approximately 5Mb to 200 Mb. The large chromosomes are referred to as macrochromosomes while the smaller chromosomes are referred to as microchromosomes. Microchromosomes account for around 30% of the chicken genome (Smith and Burt, 1998) and have previously been found to be GC rich (Auer et al., 1987). There is a large amount of variation in chromosome size, so it can difficult to define an obvious cut off point between macrochromosomes and microchromosomes. A study from 2000 suggests that there are 6 pairs of macrochromosomes and 33 pairs of microchromosomes (Smith et al., 2000). However, this was later reclassified to 5 macrochromosomes, 5 intermediate, 28 microchromosomes and the sex chromosomes (Burt, 2005). Additionally, chickens have female heterogamety as opposed to male heterogamety found in mammals.

A number of genome assemblies have been created with the most recent assembly, Gallus gallus 4.0 (Galgal4), released in 2011. Previous assemblies include WASHUC2 (Galgal3) released in 2006 and Gallus gallus 2.1. Eight of the microchromosomes remain completely absent from the current assembly, which may be due to their small size and unusually large amounts of simple repetitive sequences (Matzke et al., 1990, Fillon, 1998). Others, such as chromosome 16, are extremely under-represented, due to the presence of the highly variable major histocompatibility complex (Delany et al., 2009, Solinhac et al., 2010). A number of SNP arrays are also available for this species, including 12k, 42k and 60k chip. The most recent high density chip contains over 600,000 SNPs and was created using 243 individuals from 24 lines (Kranis et al., 2013).

This includes six commercial white egg layer lines, five commercial brown egg layer lines, four commercial broiler lines, eight inbred lines and one unselected layer line.

1.4.2 Selection Signature studies in modern chickens

A large number of studies investigating selection signatures in the modern domestic chicken have been carried out using a variety of methods. Rubin et al (2010) used a z-transformed heterozygosity (ZH_P) method to detect regions of low diversity in pooled chicken populations, including four broiler populations, four layer populations and a pool of red junglefowl. To prove that this method could find established selective sweeps, in was tested on the yellow skin allele of the *BCDO2* locus (Eriksson et al, 2008). A ZH_P score of -8.2 was found and the region exactly overlapped the previously found sweep. Three regions were found with a more extreme ZH_P score than the *BCDO2* locus. The first is a region on chromosome 1 upstream of *SEMA3A*, which encodes semaphorin 3A, a molecule with a vital function in brain development. The second was found in a non coding region upstream of the transmembrane- domain-containing protein 2A gene (*VSTM2A*). The third region also had the lowest ZH_p score (-9.2) and was located on chromosome 5 at the locus which encodes thyroid stimulating hormone receptor (*TSHR*). This receptor is known to have roles in reproduction and metabolic regulation and therefore was investigated further. A non-conservative amino acid substitution was found at residue 558, where arginine replaced glycine. Glycine at this position is conserved among all known vertebrate *TSHR* sequences and therefore is a candidate mutation for the region. This may be related to the change to non-seasonal

reproduction found in domestic animals compared to the strict seasonal reproduction in wild animals (Hanon et al, 2008).

Several other candidate genes were found using this ZH_P technique. A region on chromosome 1 was found which contains both *IGF1*, which is involved in growth and *PMCH*, which has roles in metabolic regulation and appetite. Another sweep was found at the *TBC1D1* which is fixed in both the high and low growth lines. This suggests this sweep occurred in the early development of broiler lines as the lines have been separate since 1957. A loss of function mutation in this gene has been shown to be associated with leanness in mice (Chadt et al, 2008). Finally, a novel deletion was found in the gene *SH3RF2* which was fixed in the high growth line and is found within a QTL for body weight.

This ZH_P method was used again to detect regions of low diversity in 67 chicken breeds, including commercial broiler dam and sire lines, commercial brown and white egg layers, Chinese breeds and Dutch traditional breeds (Elferink et al, 2012). This method was also tested by identifying the known selective sweep at the *BCDO2* locus (Eriksson et al, 2008). Twenty six regions were found which display strong evidence of selection, three of which were found exclusively in the broiler breed group. Candidate genes in these regions include *HNF4G*, which is associated with a higher bodyweight in knockout mice compared to the wild type (Gerdin et al, 2006) and NELL1 which is involved in bone tissue formation. Thirteen regions were also found by Rubin et al (2010) including the regions containing candidate genes *IGF1* and *PMCH*. However, the *TSHR* region was not found. This may be because only one SNP was located within the 40kb TSHR

region. This SNP was fixed in most domesticated breeds but did not reach significance as other markers in the window were segregating at high frequencies.

The ZH_P technique was also used in combination with other methods in two broiler lines divergently selected for abdominal fat content genotyped on an Illumina 60k SNP chip. (Zhang et al, 2012a). This is unusual, as the ZH_P method was previously used only on pooled lines of chickens (Rubin et al 2010, Elferink et al 2012). ZH_P and long range allele frequency differences (AFD) were used to initially identify regions under selection. Additional tests, including linkage disequilibrium (LD), allele frequency differences between the lean and fat lines, haplotype analysis using haploview (Barrett et al, 2005) and extended haplotype homozygosity (EHH) analysis using Sweep 1.1 (Sabeti et al., 2002) were then used as additional evidence for selection. The most significant region was found on chromosome Z with the largest allele frequency differences between the lean and fat lines in all chromosomes. The LD and haplotype analysis revealed two main haplotypes of which the main haplotype in the lean line did not exist in the fat line. The main fat line haplotype was found at low frequencies in the lean line. Finally, the EHH result showed that 0.73Mb of this region may have been subject to selection. The gene *PC1* (or *PCSK1*) was found in this region, which has a role in regulating insulin biosynthesis and is associated with human obesity (Choquet et al, 2011).

Nine other candidate regions were found by this study. This includes a low diversity region on chromosome 1 which contains the gene *IGF1*, which is associated with growth and fatness in chickens. Another region on chromosome 1 was identified by large allele

frequency differences and contains the *TRPC4* gene. There was stronger EHH in the fat line, suggesting that this region was due to selection on *TRPC4* in the fat line. Finally, a region on chromosome 5 was identified by large allele frequency differences, as well as significant EHH results in the lean line. Two haplotype blocks were found in the lean line, which both overlap with *NOVA1*, suggesting that this gene is involved with low abdominal fat content.

A separate study used the same two lines divergently selected for abdominal fat content, but used different statistical techniques to identify selection signatures, specifically pairwise r^2. extended haplotype homozygosity (EHH) and relative extended haplotype homozygosity (REHH) which corrects for variability in recombination rates in the EHH test (Zhang et al, 2012b). 51 positively selected core regions were found in the lean line and 57 were found in the fat line. Selection signatures were not uniformly distributed across the genome, with a large overrepresentation found on chromosomes 1 to 4. Seven genes were found in regions with significant REHH P-values ($P<0.01$) which also overlapped with previously reported QTL regions for abdominal fat (Hu et al., 2013). These include the *BBS7* gene which is related to Bardet-Biedl syndrome (BBS), a genetic disorder characterised by obesity and developmental delay (Sheffield et al, 2004). Other genes include *MAOA* and *MOAB* which have been implicated in human obesity (Need et al, 2006) and *LRP2BP* and *LRP1B* which is associated with regulation of lipid metabolism and transport of nutrients and vitamins (May et al, 2007).

Signatures of selection have also been investigated when sequencing chicken species. Two domestic chickens, a male silkie and a male TCC L2, were sequenced at 23 fold

and 25 fold average coverage using Illumina sequencing technology (Fan et al, 2013). These reads were mapped onto the chicken genome and approximately 7.6 million SNPs were identified, 42% of which had not been identified in other chicken studies before. However, not all of these may be reliable as only two chickens from two populations were sequenced. Additionally, signatures of selection were investigated within these two chickens, by detecting local reductions of variation within sliding windows and examining differences between homozygous SNP sites in the domestic lines compared to the red jungle fowl ancestor. A total of 509 genes were found which may have been under selection in the domestication process. 46 of these genes were also found by Rubin et al (2010), including *TSHR* and *IGF1*. Additionally, the role of loss of function mutations and protein coding changes in the domestication process were explored. Little evidence for selection of loss of function mutations during chicken domestication was found. 50.29% of genes within the selective sweeps were found to have protein-coding changes, compared to approximately 42% when all genes are examined, so protein-coding changes within genes may be important in the domestication of these species.

Finally, a recent paper investigated signatures of selection within commercial egg layers using 1 million SNPs (Gholami et al, 2014). The 1 million SNPs were obtained by combining three 600k SNP chips and removing duplicates SNP locations. Pairwise Wright's F_{ST} was calculated between three breeds of commercial egg layers and 14 non-commercial fancy breeds. The commercial breeds include a white egg layer breed based on White Leghorns and two brown egg layer breeds based on White Rock and Rhode Island Red. Several genes were found in sweep regions between the brown and white

layers, including *CAPN1*, which is associated with meat quality (Zhang et al, 2008) and *IGF2*, which is a fetal growth factor and is associated with carcass traits and growth in chickens (Tang et al, 2010, Jiao et al, 2013). Genes associated with both meat quality and egg production reflects that the brown egg layers were a dual purpose breed, originally bred for both meat and egg production (Muir et al, 2008). Other genes were found in the comparison between the commercial layers and the non-commercial group include *NCOA1*, which is related to total egg production and age at first egg (Huang et al, 2011) and *SPP1*, which is associated with body weight at 5 weeks (Arazi et al, 2009).

These studies have covered a number of different statistical methods in a variety of breeds. However, there are several ways for this research to be expanded upon. The majority of studies investigating local reductions in variation use SNPs chips up to a 50K density. The 600K SNP chip has been recently developed and released (Kranis et al, 2014). Using this high density SNP chip to investigate reductions in variation will allow signatures of selection to be identified with greater accuracy and therefore narrow down the possible list of genes responsible. Additionally, many studies have compared broilers and layers or commercial and non-commercial lines. Very few studies compare broiler lines to other broiler lines. SNP data from nine lines of commercial broilers is available (Andreescu et al, 2007). Calculating population differentiation between the lines may allow insight into more recent selection between broilers which occurred after the lines were split.

1.5 Aims and Objectives

The aim of this thesis is to investigate the effects of artificial selection on the genome of domesticated chickens. Regions in the chicken genome showing signs of selection are likely to be associated with commercial traits. By applying selection based tests to a number of SNP datasets characterized from nine Aviagen broiler lines (Andreescu et al., 2007) and layer lines, we may uncover genes essential to the broiler and layer lines we see today.

The subsequent paragraphs describe the contents and objectives of each chapter.

Chapter 2 investigates population differentiation in nine lines of broiler chickens genotyped on a 12k SNP chip. Commercial chicken studies often compare broiler and layer lines. However, investigating differences between broiler lines which have different selection criteria may reveal new candidate selected regions. A chromosome bound, circular permutation method was developed in order to identify highly differentiation regions and eliminate bias towards regions of low SNP density. Candidate genes and QTLs were investigated in each putative selected region and QTL enriched regions were identified.

Chapter 3 identifies regions of low diversity in one line, genotyped on three different SNP datasets of varying diversity taken at different time points. Additionally, regions of low diversity were investigated in the nine low density broiler lines utilised in chapter 2. Regions of low diversity can be created in the genome due to hitchhiking effect (Maynard Smith and Haigh, 1974) and may indicate selection has taken place. In order to confirm the presence of selection, regions displaying significant asymptotic regression

were also identified in the 600k dataset, which tests the fit of the real data to the expected model (Wiener and Pong-Wong, 2011). Regions displaying multiple signatures of selection in different datasets were investigated, as well as those which overlapped with high differentiation regions from the 12k dataset (chapter 2). Candidate genes and QTLs were investigated in each region and QTL enriched regions were identified.

Chapter 4 identifies regions of high population differentiation between a broiler line and a layer line genotyped on a 42k SNP chip. Regions of low diversity in the layer line were also investigated in this chapter. Low diversity broiler line regions were identified in chapter 3. Differences between broiler and layers may indicate selection which occurred during breed separation. To further investigate this, regions which were differentiated in both broilers and layers and between broiler lines (chapter 2) were identified. Candidate genes within regions were identified as well as QTL relating to broiler or layer traits.

Chapter 5 presents an overall conclusion of the thesis. Regions displaying multiple signatures of selection and candidate genes are discussed. The relevance of these findings to the overall field is given and further studies are proposed to build on this work.

CHAPTER TWO

Detecting Signatures of Selection in Nine Distinct Lines of Broiler Chickens

2.1 Introduction

QTL mapping has previously been utilised to identify genomic regions that influence traits of economic or biological interest in domesticated animals, but this approach requires extensive phenotypic information (Haley, 1995). A complementary approach is to use purely genotypic information to find regions of the genome showing signatures of selection, which may also be associated with traits of interest. This approach has been applied to the analysis of pigs (Rubin et al., 2012, Ai et al., 2013, Wilkinson et al., 2013), cattle (Gibbs et al., 2009, Stella et al., 2010), chickens (Rubin et al., 2010, Elferink et al., 2012) and sheep (Kijas et al., 2012, Moioli et al., 2013, Gutiérrez-Gil et al., 2014). Modern chickens (*Gallus gallus domesticus*) were domesticated from Red Junglefowl before 6000 B.C. (West and Zhou, 1989) and bred primarily for cock fighting and food. Within the last 70 years, chickens used for food were split into separate lines specialised for specific functions: broilers (meat production) or layers (egg production) (Muir et al., 2008). Both types of lines have been selected for a wide variety of traits, including welfare and disease resistance traits. However, the main broiler production traits include body weight, growth, feed conversion ratio and fat content, while layer selection is focused on reproduction-related traits. This has led to dramatic phenotypic changes, so that the genomic signatures of this selection may be detectable using statistical techniques.

Selection has a number of effects on the genome. Positive selection can cause an advantageous allele to spread throughout the population and become fixed. This causes a reduction of diversity in the population at the selected site. Additionally, genetic

variation is reduced at neutral sites linked to those under positive selection by the hitchhiking effect, which results in areas of low variation around the selected gene (Maynard Smith and Haigh, 1974). Another effect can occur between populations. Populations with limited gene flow between them and which experience different selection pressures can undergo genetic differentiation (Wright, 1951). Population differentiation measured by differences in allele frequencies between populations can be quantified by the statistic F_{ST} (Wright, 1951). When there is no selection, F_{ST} is influenced only by genetic drift, which will affect all loci in a similar way. However, when differentiation has occurred due to positive selection at particular loci across populations, an increase in F_{ST} at markers linked to the targets of selection should be observed (Akey et al., 2002).

In earlier applications of F_{ST} based tests, data simulated under neutral conditions was used to compare with real data (Beaumont and Nichols, 1996). However, recent increases in marker density allow another approach. Loci which are affected by selection show up as outliers in the tails of the empirical distribution of F_{ST}, which cuts out the need for simulations (Akey et al., 2002). This technique has been used to suggest candidate genes associated with skin wrinkling in Shar-Pei dogs (Akey et al., 2010) and coat patterns in pigs (Wilkinson et al., 2013). In dairy cattle, differentiation tests have been applied to look for genes under selection across breeds including Holstein, Normande and Montbéliarde (Flori et al., 2009). Genes showing evidence of differentiation include the *Growth Hormone* gene (*GH*) and *Insulin Growth Factor* gene (*IGF1*), which are both important in milk production.

In this study, we examined genetic differentiation between nine distinct broiler chicken lines that have been under selection for 40 generations (Andreescu et al., 2007), using genome-wide data from 12K SNPs in order to identify genomic regions where selection may have taken place during broiler specialisation. Genetic divergence between broiler lines has previously been identified (Granevitze et al., 2009), so that we can expect to detect between-line differentiation. Increased population differentiation in a specific genomic region may indicate a recent selective sweep, as sweeps reduce local variability, which also increases F_{ST} (Charlesworth, 1998, Cruickshank and Hahn, 2014).

2.2 Materials and Methods

2.2.1 Animals

The birds used in this study were from nine broiler chicken lines provided by Aviagen, a company primarily involved in broiler breeding. They originated from Plymouth Rock and Cornish lines and have been selected for at least 40 generations for a variety of traits, including growth, feed conversion ratio, reproduction and welfare characteristics. Each of these nine lines has slightly different selection criteria, allowing the production of hybrid broiler lines with different characteristics. Genotypes were available for approximately 200 individuals per line (Table 1).

Table 1: Sample sizes (numbers of individuals) of the broiler lines before and after quality control.

	Original Sample size	Sample Size after Quality Control
Line1	200	65
Line2	200	70
Line3	200	70
Line4	201	72
Line5	202	67
Line6	200	68
Line7	200	71
Line8	200	71
Line9	200	62

2.2.2 Data

These lines were genotyped for a total of 12,046 SNPs (hereafter referred to as the 12K dataset) distributed across the genome. These SNPs are a subset of the 2.8 million SNPs identified in the chicken genome sequencing project (Wong et al., 2004) and include the ~6K markers genotyped by Andreescu *et al.* (2007). The mean distance between markers was 0.13 Mb, with a standard deviation of 0.31 Mb (Powell et al., 2012). The markers are well distributed across the genome. However, there were several areas with unusually high or low SNP density; for example, an area on chromosome 3 (110-113Mb) had a particularly high SNP density while chromosome 16 had very low marker coverage, due to the presence of the highly variable major histocompatibility complex. Additionally, several low-density areas were found on chromosomes 1, 2 and 3. 11,988 of the markers had known chromosome locations and were distributed across the 28 autosomes and the Z chromosome. The number of SNPs per chromosome ranged from 27 to 2282, mainly due to the large difference in size between the macrochromosomes and microchromosomes in the avian genome. All genotyped birds were male. Quality

control procedures were used to remove individual animals from the analysis. Closely related individuals, including full sibs, half-sibs and parents were discarded. The first family member found in the pedigree file was retained for analysis and the rest were removed. This removed a total of 1169 birds from all nine lines. Additionally, 18 individuals with greater than 10% missing SNP calls were removed. This left an average of 68 individuals per line (Table 1). Quality control procedures then removed SNPs that were missing in >10% of the birds (283). As is usual in selection signature studies, SNPs located on the Z chromosome were also removed (602) from the dataset, leaving 11,003 SNPs for analysis.

2.2.3 Statistical Analysis - Population Differentiation

To investigate population differentiation, the pairwise Weir and Cockerham's F_{ST} estimator was calculated for each SNP for each pair of lines (total of 36 pairs) (Akey et al., 2002, Weir and Cockerham, 1984). In order to calculate an overall differentiation level for a given line, all pairwise F_{ST} estimates for the line were averaged (8 values per line). The individual pairwise F_{ST} values for each pair of lines were also averaged across all SNPs, yielding a genome-wide average differentiation value for each pair of lines. This allowed us to investigate the overall relatedness between the lines.

2.2.4 Sliding Windows

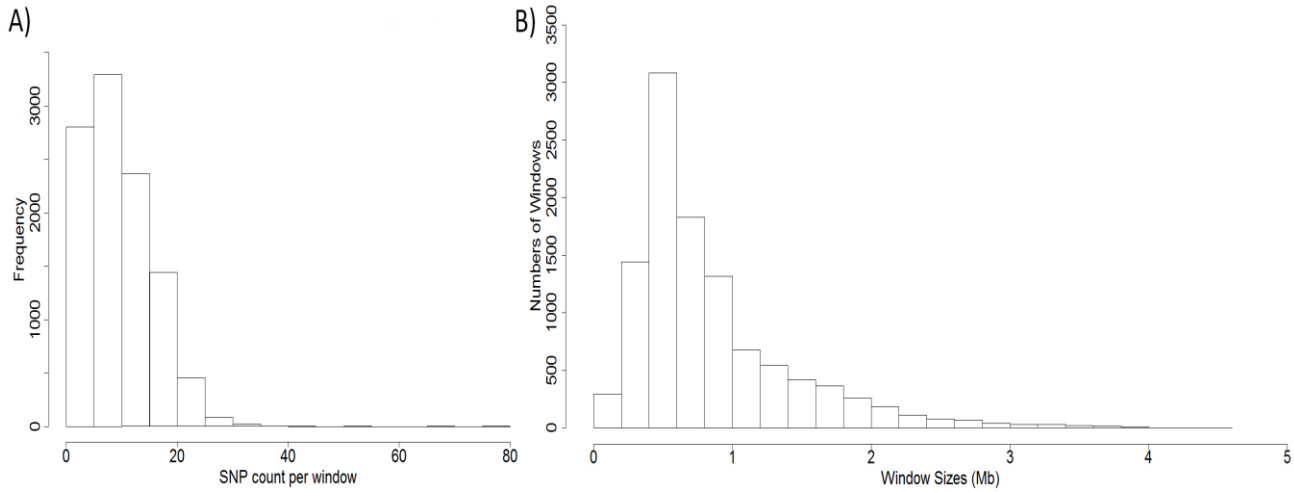
The F_{ST} results were averaged into overlapping sliding windows of SNPs, in order to smooth out stochastic variation in estimates across markers (Weir et al., 2005). Two methods were evaluated for the implementation of sliding windows across

chromosomes. The first was based on a fixed number of SNPs per window. The second was based on a fixed physical size of each window. The "fixed SNP number" method used a window with a constant number of SNPs, which moved along each chromosome one SNP at a time. The position of the window was determined by the central SNP. Several window sizes were investigated including 5, 11, 15, 21 and 31 SNPs. A window size of 11 SNPs was chosen, as this balanced the identification of signal and removal of stochastic effects. Larger window sizes diluted the signals of selection, while smaller window sizes showed too much variation, making it difficult to distinguish regions of high differentiation from random fluctuations. In this method, the size of the window changed along the chromosome, depending on the density of the SNPs in a genomic region. For the "fixed size" window method, we used a window size of 840kbp across each chromosome. The central positions of the windows were spaced 85kbp apart. This size is approximately the average size and increment used for the fixed SNP method with 11 SNP windows. The distribution of the number of SNPs in a window was investigated. To prevent windows with zero or one SNP from overly influencing results, windows with fewer than two SNPs were discarded. Some windows were identical to the previous window, as no new SNPs had been added or lost from the window. New windows which were not identical were referred to as "unique windows". This method is slightly different to the Randhawa et al. (2014) sliding window method but generates a similar number of windows, assuming the same window size (results not shown).

The distribution of SNP numbers per window based on the fixed window size method (Figure 1a) had a smaller tail than the distribution of window sizes per window using the

fixed SNP number method (Figure 1b), with the majority of windows containing fewer than 30 SNPs and an average of 10.2 SNPs per window. Additionally, some of the window sizes used in the fixed SNP number method were very large (Figure 1b), while fixed size windows with many SNPs were relatively rare (Figure 1a) and occurred mostly in one genomic region on chromosome 3. We therefore used the fixed window size method for selection analysis. The sliding window average of the pairwise average F_{ST} is referred to as the "F_{ST}-window" value.

Figure 1: Comparison of fixed window size and fixed SNP number approaches for sliding windows implementation. a: Distribution of number of SNPs for fixed size windows. b: Distribution of sizes of windows for fixed SNP number per window.



2.2.5 Signatures of Selection

Two methods were used to identify regions of selection. The first ("F_{ST}-window distribution method") identified the upper tail of the empirical distribution of F_{ST}-windows, defined as the top 0.5% of F_{ST}-window values for each line. The second method ("permutation threshold method") was used to determine a valid significance

threshold for the differentiation results. For this method, we performed circular chromosome-bound permutations (Cabrera et al., 2012, Kindt et al., 2013), using the eight pairwise F_{ST} comparisons for each SNP/line combination, calculated during the differentiation analysis for each line. This procedure (described below) maintains the order of SNPs within each chromosome, thereby preserving the internal structure of the datasets and conserving the relative distance between SNPs and linkage disequilibrium patterns across the genome. It generates an empirical distribution for the mean F_{ST} in a region for a particular line, under the null hypothesis of no overall mean differentiation, against which our observed values can be compared. After the identification of differentiated regions, we considered two separate sets of results. The first set included regions identified in multiple lines. The second set included regions unique to individual lines. In order to increase the chances of detecting genuine differentiation due to selection in both sets, only regions spread over more than one unique window were included.

2.2.6 Permutation Procedure

Each chromosome was considered to be circular and was permuted separately. The procedure implemented the following algorithm for a given chromosome and a given line: (1) SNPs were numbered based on their position on a given chromosome. (2) For that chromosome, for each of the eight pairwise comparisons associated with that line, a random number between one and the total number of SNPs in the chromosome was generated. The F_{ST} value associated with the first SNP rotated to the position of the SNP with that random number. All other F_{ST} values rotated by the same amount (the position

of any SNPs that rotated beyond the final position of the chromosome continued from the beginning of the chromosome). As a different random number was chosen for each of the pairwise comparisons, each of the 8 pairwise F_{ST} values for that SNP position were permuted by a different amount. (3) All 8 comparisons between lines were averaged into a single overall permuted differentiation value for each SNP for that line. (4) Steps (2) – (3) were repeated for all chromosomes for the current line. (5) Steps (2) – (4) were repeated for each of the other 8 lines. (6) Steps (2) – (5) were repeated 100,000 times, giving 100,000 permuted average F_{ST} values for each position/line combination.

In order to make these permuted results directly comparable to the F_{ST}-window distribution analysis, the average F_{ST} values for each position were again averaged into overlapping sliding windows. These sliding windows were 840kb long and spaced 85kb along the genome, as in the F_{ST} windows analysis. Windows with fewer than two SNPs per window were again discarded. This resulted in a total of 10,498 overlapping windows in both the real data and permutation analysis. These 10,498 overlapping windows are equivalent to 1076 non-overlapping, and hence independent, windows. We thus desired to set our Bonferroni corrected threshold significance at $p = 0.05/1000$ to take the number of independent windows into account, giving a p of approximately 0.00005 for each line. From the permutations, we obtained 100,000 average F_{ST} values for each window. The fifth largest average F_{ST} value for each window was used to set the individual empirical threshold for that window, providing a 5/100000 = 0.00005 threshold for each window.

"Regions" were defined by manually looking for significant windows adjacent to each other. A region could include a small gap of one or two windows to account for windows that were discarded for including fewer than 2 SNPs. Even with this gap, there is still a large overlap between adjacent windows, due to the markers shared between adjacent windows in a sliding window approach. Regions can overlap if there are small (but greater than two windows) gaps. These regions were then compared for overlaps between the lines.

2.2.7 QTLs and genes within differentiated regions

In order to investigate QTLs previously discovered by other studies, we probed the Animal QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/index) (Hu et al., 2013). The database includes a total of 3,811 QTLs in chickens, representing 297 traits from 191 publications, and the base pair positions reported are on the Galgal4 build of the chicken genome. As our data were recorded on the WASHUC2 build of the chicken genome, we used UCSC batch coordinate conversion (*liftover*) (http://genome.ucsc.edu/cgi-bin/hgLiftOver) with default settings to position the regions on Galgal4. The Chicken QTL database represents the coordinates of QTLs as a map property known as “span”. Span represents a genomic area but does not incorporate statistical confidence. Many QTL had very large spans, and no peak position was available in the database. In order to narrow down the possible QTL for each region, we assigned a peak position as the midpoint of the QTL span. We searched for the regions identified in this database and included any QTLs whose assigned peak position fell within these regions. "Broiler QTLs" were defined as those meeting certain criteria: (1)

they were assigned as production QTLs in the Chicken QTL DB and (2) they were directly related to meat production. For example, growth, feed conversion ratio, body weight and breast muscle weight are all considered to be broiler QTL. A full list of broiler traits is available in Table S1.

Additionally, the genes in the regions showing high differentiation were investigated using Ensembl/Biomart, using the dataset *Gallus gallus* genes (Galgal4) from the database Ensembl genes 74. (http://www.ensembl.org/biomart). The number of genes in the region and their functions (if available) were recorded.

2.2.8 Significance of QTL Results

In order to test if certain regions were enriched for broiler QTLs, we applied the following approach: for each genomic region that exceeded the permutation derived significance threshold, a region of that size was randomly assigned 100,000 times to positions on the chicken genome. The number of broiler QTL peak positions present in these simulated regions, based on the Animal QTL Database, was calculated for each iteration. The top 5% of these 100,000 QTL counts were determined for each region size. If the actual number of QTL for a region exceeded its top 5% threshold, it was classed as enriched.

Additionally, to determine whether there were more differentiated regions that contained one or more QTL than would be expected by chance, we calculated the total QTL coverage of the genome to assess the proportion of QTL across the genome that were found within an average-sized differentiation region. To calculate this, we moved along

the chicken genome one base pair at a time and recorded whether each position was within 1.599Mb (the average size of a differentiated region) of a QTL peak position, such that if the differentiation region covered that position, it would also include a QTL position ("overlap position"). The number of base pairs across the genome which were overlap positions was calculated, and divided by the total number of base pairs in the genome. To calculate the expected number of differentiated regions containing QTL, this proportion was multiplied by the total number of differentiated regions.

2.3 Results

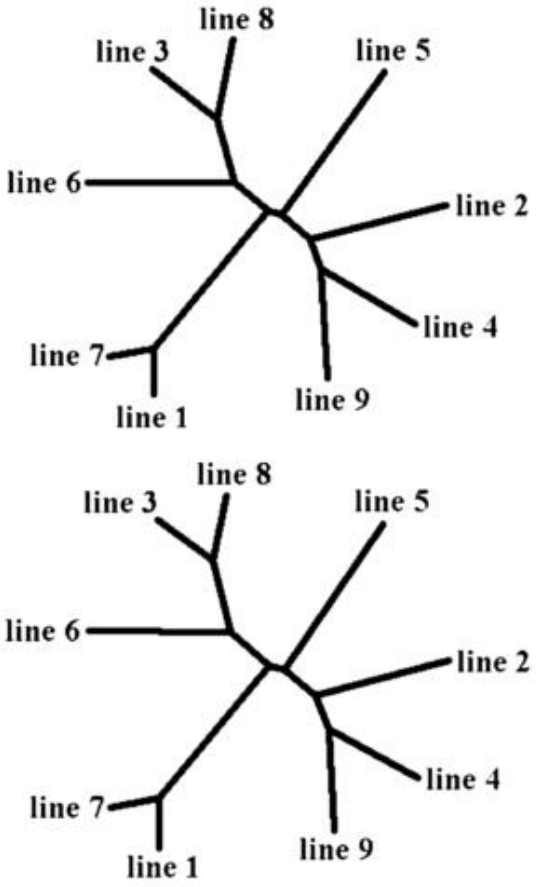
2.3.1 Relationships between Lines

The genome-wide differentiation (mean F_{ST}) values for several pairs of lines were much lower than others (Table 2). Lines 1 and 7 have an extremely low differentiation value at 0.015, suggesting they are more closely related than other lines. Lines 2, 4 and 9 appear to be closely related. Lines 3, 6 and 8 also appear to be related, but not as closely as the previous groups. Line 5 does not seem to be closely related to any of the other eight lines. These results are consistent with a tree generated in a previous study with the same lines, based on allele frequencies on chromosomes 1 and 4 (Andreescu et al., 2007) – see Figure 2.

Table 2: The average differentiation values for each pair of lines, with standard errors included in parentheses. Closely related lines, such as lines 1 and 7, have much lower differentiation than lines with no shared ancestry (Andreescu *et al.*, 2007) (Figure 2).

Line	1	2	3	4	5	6	7	8	9
1	X	0.174 (0.0019)	0.178 (0.0021)	0.171 (0.0019)	0.153 (0.0018)	0.160 (0.0019)	0.015 (0.0003)	0.173 (0.0020)	0.174 (0.0019)
2		X	0.162 (0.0019)	0.091 (0.0012)	0.142 (0.0017)	0.158 (0.0018)	0.170 (0.0019)	0.155 (0.0018)	0.093 (0.0012)
3			X	0.153 (0.0017)	0.156 (0.0018)	0.109 (0.0014)	0.173 (0.0020)	0.035 (0.0006)	0.156 (0.0018)
4				X	0.134 (0.0016)	0.148 (0.0017)	0.167 (0.0018)	0.147 (0.0017)	0.056 (0.0008)
5					X	0.134 (0.0016)	0.147 (0.0017)	0.150 (0.0017)	0.137 (0.0016)
6						X	0.155 (0.0018)	0.103 (0.0013)	0.150 (0.0017)
7							X	0.168 (0.0020)	0.171 (0.0019)
8								X	0.146 (0.0017)
9									X

Figure 2: Phylogenetic trees based on marker allele frequencies for chromosome 1 and chromosome 4, obtained using the UPGMA algorithm. Originally produced in Andreescu *et al.* (2007).



2.3.2.1 Signatures of Selection found in Multiple Lines

Multiple signatures of selection were identified. The F_{ST}-window distribution analysis identified 87 regions as outliers in terms of their pairwise differentiation levels across the nine broiler lines. Of these regions, 14 were shared between at least 2 lines and spread over more than one unique window, on chromosomes 1, 2, 3, 4, 5, 11 and 14. These 14 regions were positioned on the Galgal4 genome build using *liftover*. The permutation threshold method found a total of 214 regions. Of these regions, 51 were

shared between at least 2 lines and spread over more than one unique window, on chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 18, 20, 21, 26, 27 and 28 (Table 3). Regions 15 and 16 and regions 24 and 25 were found to overlap, due to the method used for defining regions and the use of overlapping sliding windows. Twelve of the 14 F_{ST}-window regions were also found in the permutation threshold regions. The other two regions were found in fewer lines using the permutation threshold method; because they were only identified in a single line, they feature in Table 4 (see below) instead of Table 3. A region at 5: 2.52Mb - 4.04Mb was found in lines 4 and 6 in the F_{ST}-window distribution results, but only in line 6 using the permuted thresholds. Another region was found at 11: 1.89Mb - 2.65Mb in lines 4, 7 and 8 by the F_{ST}-window distribution method, but only in line 7 in the permutation threshold results. The other 12 regions were present in both sets of results, but some additional lines were gained from the permuted threshold method. For example, a region at 1:53.51Mb - 54.71Mb was found for lines 4 and 9 in the F_{ST}-window distribution results but the overlapping Region 2 from the permutation threshold results showed high differentiation in lines 1, 2, 4, 5, 6, 7 and 9.

Table 3: Regions displaying putative signatures of selection that were found in multiple lines. The criteria for inclusion in this table are as follows: windows in the region exceeded their permuted significance threshold, the region was found in more than one line and spread over more than one unique window. The 95% significance threshold is included for the five regions for which the number of QTL exceeded that expected by chance. Regions also found by the F_{ST}-window distribution method are highlighted in bold.

Region No	Chromosome	Start (Mb)	End (Mb)	Size (Mb)	Lines	QTLs	Broiler QTL	Genes	95% Threshold for Broiler QTL number
1	1	2.74	4.52	1.77	3, 5, 8	0	0	24	
2	**1**	**52.13**	**54.71**	**2.59**	**2, 4, 5, 6, 7, 9**	**17**	**5**	**34**	
3	1	102.21	103.65	1.43	1, 4, 7	1	1	13	
4	1	130.48	132.07	1.59	1, 7	2	1	27	
5	1	133.97	136.13	2.16	2, 3, 4, 8, 9	1	1	21	
6	**1**	**140.67**	**142.08**	**1.41**	**4, 6, 8, 9**	**8**	**3**	**6**	
7	1	145.22	146.81	1.59	2, 8	3	1	9	
8	**1**	**177.11**	**178.95**	**1.85**	**2, 3, 8**	**1**	**1**	**27**	
9	1	187.21	188.28	1.08	2, 9	5	2	12	
10	**2**	**16.71**	**19.49**	**2.78**	**2, 3, 4, 5, 7, 9**	**2**	**1**	**34**	
11	2	24.14	25.98	1.85	1, 4, 5, 7, 9	0	0	19	
12	2	102.42	103.60	1.18	3, 6, 7	6	5	19	
13	2	111.56	113.96	2.41	1, 6	22	4	18	
14	2	122.56	123.89	1.33	5, 8	2	1	10	
15	2	139.95	142.13	2.18	1, 2, 4, 5, 7, 8,	1	1	20	
16	**2**	**141.88**	**143.47**	**1.59**	**1, 2, 4, 7, 9**	**0**	**0**	**4**	
17	3	14.50	16.66	2.16	1, 4, 5, 6, 7, 8,	0	0	72	
18	**3**	**34.21**	**35.66**	**1.46**	**3, 8**	**0**	**0**	**17**	
19	3	38.14	39.36	1.22	1, 6	3	0	19	
20	3	50.58	51.84	1.26	1, 5	5	0	23	
21	3	58.32	59.90	1.58	1, 9	8	1	17	
22	**4**	**17.84**	**19.14**	**1.30**	**1, 3, 6, 7**	**0**	**0**	**7**	
23	4	19.56	20.73	1.17	1, 7	0	0	18	

24	**4**	**57.62**	**59.21**	**1.59**	**2, 3, 8, 9**	**5**	**1**	**13**	
25	4	58.64	59.69	1.06	3, 8	2	1	12	
26	4	63.62	64.79	1.17	2, 8	6	4	17	
27	**4**	**84.77**	**87.19**	**2.42**	**1, 7**	**8**	**2**	**47**	
28	**5**	**27.42**	**29.09**	**1.67**	**1, 7**	**3**	**2**	**37**	
29	5	55.37	56.90	1.53	1, 2, 3, 4, 6	21	6	29	
30	6	10.48	13.11	2.64	1, 3, 4, 7, 8, 9	0	0	52	
31	7	12.03	13.78	1.75	1, 7	1	1	29	
32	7	17.52	18.93	1.41	1, 7	19	18	31	7
33	7	23.20	24.80	1.60	4, 9	5	4	10	
34	7	31.04	32.21	1.18	4, 9	0	0	4	
35	8	26.34	28.01	1.67	1, 3, 6, 7	18	9	50	9
36	9	8.27	9.78	1.51	3, 8	5	4	22	
37	10	4.66	5.91	1.25	2, 3	0	0	22	
38	11	0.00	1.37	1.37	2, 4, 9	0	0	75	
39	**11**	**13.04**	**14.79**	**1.75**	**1, 3, 7, 8**	**0**	**0**	**11**	
40	12	9.09	10.31	1.22	6, 8	1	1	21	
41	13	12.42	13.65	1.23	3, 8	13	7	41	7
42	13	15.61	17.22	1.61	2, 6	3	3	48	
43	14	0.26	1.77	1.50	2, 4	0	0	35	
44	**14**	**9.80**	**10.85**	**1.05**	**1, 7, 9**	**1**	**1**	**10**	
45	15	1.34	2.77	1.42	1, 5, 7	1	0	16	
46	18	7.41	9.00	1.59	1, 4, 7	14	9	24	9
47	20	6.17	7.34	1.17	3, 8	1	0	25	
48	21	1.30	2.47	1.17	1, 7	0	0	48	
49	26	3.43	4.61	1.18	1, 7	3	2	59	
50	27	2.72	4.50	1.78	1, 4, 5, 6, 7, 8,	46	39	114	9
51	28	0.00	1.49	1.49	1, 2, 7	0	0	81	

Thirty-seven of the permutation threshold regions were not identified using the F_{ST}-window distribution method. These include some regions that showed high differentiation across many lines: Region 17 (3:14.504Mb - 16.661Mb) showed high differentiation for all lines except line 2 and 3 and Region 50 (27: 2.720Mb - 4.502Mb) also showed high differentiation for all lines except 2 and 3. Because most of the F_{ST}-window regions were a subset of the permutation threshold regions, we chose to concentrate on the 51 permutation threshold based regions.

Several lines were consistently found to have shared significant regions; only 9 regions included lines with no shared branches in the phylogeny shown in Figure 2. Lines 1 and 7, which are on the same branch, were found to share differentiation signals in 22 regions, of which 9 were unique to lines 1 and 7 (the others were shared with additional lines). An additional shared phylogeny branch involves lines 2, 4, and 9. At least two of these lines appear together in 17 regions, 5 of which were unique to these lines. The last phylogenetic group of lines includes 3, 6 and 8; at least two of these appear together 19 times, and 6 of these regions are unique to this phylogeny branch. Finally, line 5 appears 10 times, at least once with each of the other lines (most shared with line 7 and least with line 3).

2.3.2.2 Conversion of coordinates to Galgal4 assembly

Of the 51 significantly differentiated regions, 33 were converted directly from the WASHUC2 (Galgal3) chicken genome assembly to the newer chicken genome

assembly, Galgal4, using *liftover*. Sixteen of the 18 remaining regions were converted by submitting their start and end points separately. For the other two regions (22 and 40), either the start or end points could not be successfully converted by *liftover*. In order to estimate the position of these missing coordinates, genes nearby were identified in the WASHUC2 genome build and positioned in the Galgal4 build. The closest gene to the end coordinate of region 22 is *SLITRK2.* The end coordinate of this gene in Galgal4 (19,134,524bp) was used as the missing coordinate for this region, which is likely to be a conservative estimate (i.e. a smaller region), as the 3' end of the gene is found ~0.39Mb upstream of the end of region 22 according to WASHUC2. For region 40, the closest gene in the WASHUC2 build is *coatomer subunit gamma gene (COPG).* As the start coordinate of this gene is only 912bp away from the missing coordinate, this was used as an estimate of the start position for the region (9,090,877bp according to Galgal4).

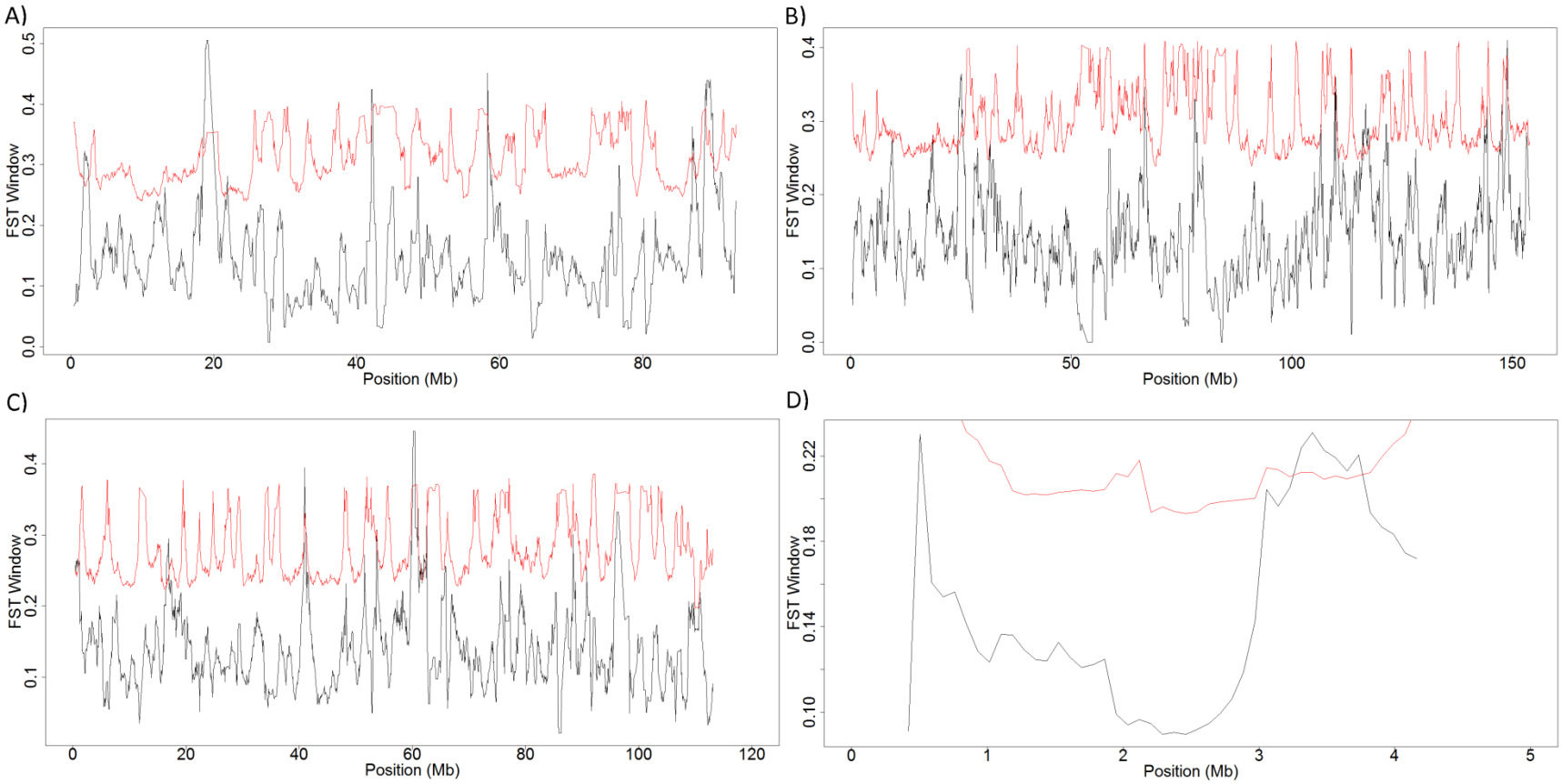
Details of all identified regions, including all QTLs and genes within the regions, can be found in Tables S2 and S3. A few selected examples of these regions are described below (with Galgal4 coordinates). These include the most highly differentiated region, the regions shared by the most lines and the largest region of physical size.

2.3.2.3 Region 22 (4: 17.84 - 19.134Mb) – the Region with the Highest Differentiation

This 1.30Mb region showed significant differentiation in four lines: 1, 3, 6 and 7. Line 1 contains the window with the highest differentiation across the genome (Figure 3a). There were no QTL found with a peak position in this region. However, several QTL

had spans which overlapped with this region, including body weight, shank growth and visceral fat weight. Shank traits were found in multiple studies. 8 genes were found in this region and 31 genes are located within 1Mb of the region including a melatonin receptor.

Figure 3: Graphs of position against F_{ST}-Window In line 1 for chromosomes a) 4, b) 2, c) 3 and d) 27. The red line represents the permuted significance threshold for each window.



There were three regions shared by seven broiler lines; these are described below.

2.3.2.4 Region 15 (2: 139.95 - 142.128Mb) – region shared by the majority of lines (1)

This 2.2Mb region showed significant differentiation in lines 1, 2, 4, 5, 7, 8 and 9 (Figure 3b). One QTL for body weight at 200 days had its peak position in this region. The spans for several other QTL reported in two studies overlapped the region; these include drumstick and thigh muscle weight, fat distribution, blood cell count, breast muscle weight and body weight (Table S2, Region 15). 20 genes were found within this region and 5 additional genes were located within 1Mb of the region (Table S3, Region 15).

2.3.2.5 Region 17 (3: 14.50 - 16.66Mb) – region shared by the majority of lines (2)

The second region present in seven lines is 2.17Mb in size, and appears in lines 1, 4, 5, 6, 7, 8 and 9 (Figure 3c). No QTL were found with a peak position in this region, although a number of QTL had spans that overlapped with this region, including body weight, abdominal fat weight, breast muscle percentage, and egg weight. Body weight and abdominal fat QTL were found in multiple studies (Table S2, Region 17).

2.3.2.6 Region 50 (27: 2.72 - 4.506Mb) – region shared by the majority of lines (3)

The final region found in seven lines is 1.78Mb in size and appears in lines 1, 4, 5, 6, 7, 8 and 9 (Figure 3d). This is a QTL- and gene-rich area, overlapping 59 QTL spans and including 107 genes (Table S3, Region 50). Of the 59 QTL spans, 46 had a peak position that fell within this region. The QTLs include traits such as body weight, carcass protein

content, growth rate, shank weight percentage and abdominal fat weight. QTL for body weight and abdominal fat weight were found by multiple studies (Table S2, Region 50).

2.3.2.7 Region 10 (2: 16.711 – 19.493Mb) – largest region in terms of physical size

A 2.78Mb region on chromosome 2 was the largest region found in the permutation threshold analysis (Figure 3b). Significant differentiation in this region was found for lines 2, 3, 4, 5, 7, and 9. One broiler QTL for feed intake had a peak position within this region. A total of 34 genes were found in this region (Table S3, Region 10).

2.3.3 Significance of QTL Results

Approximately 82% of the genome was located within 1.599Mb of any QTL. Thus with a total of 51 differentiated regions, we expect 42 regions on average to contain QTL by chance. As only 36 of the differentiated regions contained one or more QTL, we did not observe more regions containing QTL than would be expected by chance. However, five regions (32, 35, 41, 46 and 50) were found to be enriched for the numbers of QTL (the number of broiler QTL exceeded or equalled the top 5% threshold) (Table 3).

2.3.4 Line-Specific Signatures of Selection

As well as investigating regions which appeared in a minimum of two lines, regions only identified for a single line were also examined using the permutation threshold method. A total of 87 regions were found that were significantly differentiated in only one line, had a greater F_{ST} value than the permuted threshold and were found in two or more unique windows (Table 4). These regions were placed on the Galgal4 genome build, as

described above. Line 5 had the largest number of line specific region (18 regions), followed by Line 6 (16 regions). Lines 3 and 7 contained the fewest number of line specific regions, with only four each. Three regions exceeded or equalled their permuted threshold for broiler QTL. These included region 57, a 1Mb region on chromosome 6 identified in line 5; region 62, a 0.75Mb region on chromosome 9 identified in line 1 and region 69, a 1.33Mb region on chromosome 10 identified in line 5.

Table 4: Line-specific regions displaying putative signatures of selection. The criteria for inclusion in this table are as follows: windows in the region exceeded their permuted significance thresholds, the region was spread over more than one unique window and the region was unique to a single line. The 95% significance threshold is included for the three regions which exceeded the number of QTL expected by chance. Regions also found by the F_{ST}-window distribution method are highlighted in bold.

Region No	Chromosome	Start (Mb)	End (Mb)	Size (Mb)	Line	QTLs	Broiler QTL	Genes	95% Threshold for Broiler QTL number
1	1	5.36	6.37	1.01	5	0	0	14	
2	1	10.20	11.68	1.48	1	2	2	7	
3	1	12.70	13.70	1.00	6	1	0	11	
4	1	15.06	16.00	0.94	2	0	0	15	
5	1	27.70	28.93	1.22	5	5	3	8	
6	1	31.13	32.48	1.35	7	0	0	8	
7	1	34.56	36.15	1.59	8	10	10	27	
8	1	55.56	56.81	1.25	6	4	4	22	
9	1	61.33	62.68	1.35	6	3	2	26	
10	1	64.88	66.06	1.18	5	8	4	18	
11	1	73.03	74.35	1.32	2	3	3	16	
12	1	81.19	82.60	1.41	9	2	0	15	
13	1	84.71	85.88	1.17	4	4	2	11	
14	1	86.06	87.97	1.91	5	7	4	19	
15	1	107.52	109.11	1.59	6	1	1	25	
16	1	109.35	110.16	0.80	3	0	0	15	
17	1	117.45	118.37	0.92	3	1	1	13	
18	1	119.14	120.39	1.25	2	0	0	16	
19	1	122.09	123.24	1.15	8	3	1	15	
20	1	150.42	151.99	1.57	5	11	8	2	
21	1	190.49	191.81	1.32	9	0	0	2	
22	2	0.60	1.87	1.27	2	0	0	21	
23	2	3.91	5.33	1.42	6	16	9	46	

24	2	20.51	21.43	0.92	5	0	0	14	
25	2	31.04	32.24	1.19	7	2	0	23	
26	2	55.06	55.99	0.93	3	0	0	15	
27	2	65.20	66.36	1.17	2	5	0	10	
28	2	69.63	70.67	1.04	8	0	0	1	
29	2	70.90	72.63	1.73	8	1	0	4	
30	2	99.48	100.42	0.94	4	4	4	8	
31	2	106.17	107.26	1.09	1	1	1	11	
32	2	120.45	122.04	1.59	8	3	0	19	
33	2	146.68	147.93	1.25	9	0	0	7	
34	3	0.02	1.36	1.34	1	0	0	15	
35	3	0.86	1.95	1.09	2	0	0	6	
36	3	2.13	3.40	1.27	4	6	3	37	
37	3	2.98	3.91	0.93	8	11	7	27	
38	3	10.86	12.38	1.52	5	0	0	24	
39	3	12.73	14.15	1.43	5	0	0	16	
40	3	16.25	17.67	1.42	8	3	0	40	
41	3	19.96	20.88	0.92	9	8	3	7	
42	3	45.94	47.28	1.34	2	7	3	7	
43	3	48.81	49.73	0.92	5	1	1	10	
44	3	70.34	71.45	1.10	5	7	4	11	
45	3	78.55	79.54	0.99	4	0	0	15	
46	4	6.15	7.07	0.92	5	0	0	1	
47	4	22.25	23.25	1.00	3	0	0	8	
48	4	53.21	54.21	1.00	9	8	3	17	
49	4	55.93	57.11	1.18	1	0	0	15	
50	4	68.00	68.91	0.91	2	5	4	17	
51	4	80.58	81.88	1.31	5	6	3	19	
52	4	83.08	84.26	1.18	1	7	4	10	
53	5	1.34	2.42	1.08	6	0	0	34	

54	**5**	**2.52**	**4.04**	**1.53**	**6**	**0**	**0**	**21**	
55	5	29.61	31.10	1.50	6	13	3	16	
56	6	7.02	8.44	1.42	2	2	2	18	
57	6	31.13	32.13	1.00	5	13	9	22	9
58	7	22.11	23.19	1.08	4	0	0	13	
59	8	5.71	7.05	1.34	4	2	2	40	
60	8	10.68	11.76	1.08	2	8	3	26	
61	9	15.20	16.96	1.76	1	0	0	58	
62	9	16.67	17.43	0.75	1	8	8	15	7
63	9	19.93	21.02	1.09	6	0	0	15	
64	9	22.29	23.29	1.00	9	7	1	44	
65	10	1.75	2.76	1.01	2	0	0	57	
66	10	2.20	3.22	1.02	1	0	0	23	
67	10	7.68	8.63	0.94	2	0	0	15	
68	10	15.57	16.83	1.26	6	2	1	17	
69	10	17.82	19.15	1.33	5	18	11	40	11
70	**11**	**1.72**	**2.65**	**0.93**	**7**	**2**	**0**	**46**	
71	12	5.31	6.56	1.26	5	7	6	13	
72	12	8.09	9.00	0.91	4	1	1	21	
73	14	10.13	11.48	1.35	6	4	4	3	
74	15	7.17	8.07	0.90	6	6	5	36	
75	17	3.99	4.97	0.98	6	0	0	27	
76	17	9.48	10.40	0.92	7	0	0	25	
77	18	6.00	7.74	1.74	6	2	2	54	
78	19	0.00	1.23	1.23	5	5	0	51	
79	19	2.75	4.10	1.36	8	5	4	25	
80	20	0.60	2.17	1.57	2	5	5	45	
81	20	9.62	10.92	1.30	6	4	4	101	
82	20	13.22	14.14	0.91	5	0	0	21	
83	22	2.59	3.94	1.35	4	10	4	39	

84	23	3.48	5.12	1.64	9	9	5	73	
85	24	0.00	0.90	0.90	4	1	0	36	
86	24	0.66	1.66	1.00	5	0	0	22	
87	26	0.00	1.50	1.50	6	6	3	83	

Details of all identified regions, including all QTLs and genes within the regions, can be found in Tables S4 and S5. Some selected examples of these regions are described below. These include the largest region of physical size, the region containing the most QTL peak positions and the most gene rich region.

2.3.4.2 Region 14 (1: 86.06-87.97Mb) – largest region in terms of physical size

The largest region of physical size is a 1.91Mb region found on chromosome 1 in line 5. Seven QTLs, each from an independent study, were found in this region, four of which were broiler QTL. These include body weight, visceral fat weight, pectoralis minor percent and shank weight percentage (Table S4, Region 14). A total of 19 genes were found in this region (Table S5, Region 14).

2.3.4.3 Region 69 (10: 17.82-19.15Mb) – region with greatest number of QTL peak positions

The region with the greatest number of QTL peak positions, and also statistically enriched for QTL, is a 1.33Mb region on chromosome 10 found in line 5. A total of 18 QTL were found in this region, 11 of which were broiler QTL. These include seven body weight QTL, found in five independent studies, as well as QTL for abdominal fat percentage, carcass weight and growth (Table S4, Region 69). Additionally, 40 genes were found in this region (Table S5, Region 69).

2.3.4.4 Region 81 (20: 9.62-10.92Mb) – region containing the greatest number of genes

The most gene-enriched region is a 1.30Mb region on chromosome 20, which was found in line 6. There were a total of 101 genes overlapping with this region. Four broiler QTL, all identified in the same study, were found with peak positions in this region (Table S4, Region 81). These include QTL for drumstick and thigh weight, shank length and wing weight.

2.4.1 Discussion

Artificial selection has produced the distinct lines of domesticated chickens we see today and has drastically increased the efficiency of food production. We have identified possible signatures of selection between nine broiler lines by analysing genome-wide SNP data. A total of 51 regions were discovered that showed evidence of differential selection across more than one line, while 87 regions showed evidence of differential selection in a single line.

2.4.2 Methodological issues

The overlapping sliding window method used here is slightly different from the majority of sliding window methods. Usually, a fixed number of SNPs is used as the size of a window, and moves along the genome one SNP at a time (Stella et al., 2010, Wilkinson et al., 2013). However, in this dataset there were extremely variable SNP densities in different areas of the genome, which would have resulted in substantial variation in the physical size of SNP-based windows and thus invalidated the main approach of comparing average F_{ST} values across windows. The fixed-window size method applied here allows the SNP density to vary across the genome. In order to prevent the results being affected by windows with very low SNP density, any windows with fewer than two SNPs were discarded.

Previous studies have used the empirical tails of the F_{ST} distribution to identify regions of interest (Akey et al., 2002). However, due to large variation in the number of SNPs per window described above, the empirical tails of the F_{ST} distribution are more likely to include windows with few SNPs, as these will show greater variance and hence more extreme F_{ST} values. This could lead to a bias in the regions selected as significant. A circular permutation method (as suggested by Cabrera *et al.*, 2012 and Kindt *et al.*, 2013) was therefore utilised to calculate a significance threshold for every window for all nine lines, reducing this bias towards low SNP windows and also preserving the structure of the individual chromosomes, the relative distance between SNPs, and the patterns of linkage disequilibrium.

Fifty-one regions that were differentiated in at least two lines were identified by the permutation threshold method, while 14 regions were identified by the F_{ST}-window distribution method. The mean SNP count per window for the 51 permutation threshold regions found in multiple lines was 12.37 (s.e. 0.78), significantly greater than the 5.65 (s.e. 1.22) SNPs/window for the 14 regions identified by the F_{ST}-window distribution ($p<0.001$), consistent with the above prediction. The average SNP count across all windows in the genome (with at least two SNPs) fell between the average values for the regions identified by the two methods (10.16, s.e. 0.062). All of the F_{ST}-window distribution regions overlapped with regions identified using the permutation threshold prior to the imposition of additional criteria, although there were some differences with respect to the lines that were represented in a given region.

2.4.3 Composition of differentiated regions

Thirty-six of the 51 differentiated regions found in multiple lines based on the permutation threshold results overlapped QTL from the Animal QTL Database, including QTLs for broiler production traits. Region 50 on chromosome 27 is particularly interesting in that it contained peak positions for 39 broiler QTL, well in excess of the number expected by chance. This region had already been highlighted because it was highly differentiated in seven out of the nine lines, suggesting selection across broiler lines for production-related characteristics. In this region, lines 1, 4, 7 and 9 showed the greatest allele frequency differences from other lines and levels of differentiation substantially greater than their genome-wide averages. The related pairs of lines (1 and 7, 4 and 9) show similar allele frequencies (Table 2). This suggests that the main selection in this region has been on this group of lines.

The QTLs in region 50 were reported in 12 independent studies of broiler production traits which included body weight at various days (Ambo et al., 2009, Podisi et al., 2011), abdominal fat weight percentage (Campos et al., 2009, Ankra-Badu et al., 2010), shank weight percentage (Baron et al., 2011), drumstick and thigh weight percentage (Ankra-Badu et al., 2010) and carcass fat and protein content (Nones et al., 2012) (Table S2, Region 50). Two insulin-like growth factor (IGF) binding protein genes are found in this region: *insulin-like growth factor 2 mRNA binding protein 1* (*IGF2BP1*), which binds to the mRNA of *insulin-like growth factor 2 (IGF2)* and other genes and regulates their translation, and *insulin-like growth factor binding protein 4 (IGFBP4),* which binds IGFs and alters their interaction with cell surface receptors. Furthermore, there are several genes located within 1 Mb of Region 50 that have been suggested to account for broiler QTL, including *growth hormone 1 (GH1)* and *corticotropin releasing hormone receptor 1 (CRHR1)* (Nones

et al., 2012). Our findings suggest that one or more of these growth-related genes may have been differentially selected in the nine broiler lines. Insulin-like growth factors are proteins similar in structure to insulin whose production is stimulated by growth hormone (GH). These proteins have been shown to have an important role in growth. For example, mice with a mutation in *IGF1* are approximately 30% of normal size and infertile (Baker et al., 1996).

Two other genes of note were found in the differentiated regions detected in multiple lines. *LEPR (leptin receptor gene*, Region 35) is the receptor for the leptin hormone, which is secreted by adipose tissue and plays a key role in appetite and energy expenditure in mammals. SNPS in *LEPR* were previously found to differ in allele frequencies between broilers and layers (Twito et al., 2011). *MXD4 (MAX Dimerization Protein 4*, Region 27) is a tumor suppressor gene and is involved in cell growth. It was recently identified in a GWAS study as a positional candidate gene associated with abdominal fat weight in chickens (Sun et al., 2013).

The majority of the 51 differentiated regions found in multiple lines also overlapped with regions showing evidence of selection in one or more previous studies of chickens (Rubin et al., 2010, Zhang et al., 2012a, Zhang et al., 2012b, Fan et al., 2013). These include Regions 15-16, the first of which was mentioned above because it was shared by seven of the nine broiler lines. These regions overlapped regions of high homozygosity in a set of domestic lines and a subset of commercial broilers (Rubin et al., 2010) and in two Asian breeds (Fan et al., 2013). However, no functional candidate genes were found within or within 1Mb of this region. Region 47 was also notable in that it overlapped with regions identified in three independent studies: regions of high homozygosity in commercial broilers (Rubin et al., 2010)

and two Asian breeds (Fan et al., 2013) and a region showing allele frequency differences between lines under divergent selection for abdominal fat (Zhang et al., 2012a). However, this region contained no functional candidate genes. A *growth hormone releasing hormone precursor* gene was found within 1Mb of this region which is related to growth hormone. As observed for region 50, the primary allele frequency differences in most of the other QTL-enriched regions were observed between lines 1 and 7 and lines 4 and 9.

Fifty-six of the 87 regions that were differentiated in individual lines overlapped with QTL from the Animal QTL Database (Table S4). Of these 56 regions, 47 contain QTL associated with broiler production traits. Several of these regions also overlap with regions showing evidence of selection in previous chicken studies. Region 87 (line 6) on chromosome 26 overlaps with a region displaying high linkage disequilibrium values in two lines under divergent selection for abdominal fat (Zhang et al., 2012a) and a region of high homozygosity in commercial broilers (Rubin et al., 2010). Regions 30 (line 4), 37 (line 8), 52 (line 1), 61 (line 1) and 81 (line 6) overlap with regions displaying significant relative extended haplotype homozgosity (REHH) in the lines divergently selected for abdominal fat (Zhang et al., 2012b). Line 5 contains the largest number of line-specific differentiated regions. This is consistent with the overall line relationships (Table 2, Figure 2), which show that line 5 is not closely related to any of the other lines.

A number of candidate genes known to be related to growth were found in the line-specific differentiated regions. *Growth hormone releasing hormone receptor* (*GHRHR*, region 22, line 2) is the receptor the binds growth hormone-releasing hormone which causes the production and release of growth hormone. Mutations in

this gene have been known to cause isolated growth-hormone deficiency, which is associated with small size in mice (Godfrey et al., 1993). *Insulin-like growth factor 1 receptor* (*IGF1R*, region 68, line 6) is activated by insulin-like growth factor 1, which is important in growth. Mice lacking the *IGF-1* receptor die during development and have massively reduced body mass compared to normal mice (Liu et al., 1993). Finally, *insulin-like growth factor 2 mRNA binding protein 3* (*IGF2BP3*, region 25, line 7) has been known to repress translation of *insulin-like growth factor II*, a growth-promoting hormone, during late development (Nielsen et al., 1999).

The majority of the above-mentioned QTL studies investigated broiler-layer crosses. Therefore, these QTLs represent differentiation between broiler and layer lines but not necessarily between different broiler lines. The absence of broiler QTL for some of the regions identified in these studies may reflect this bias towards regions differentiated between broilers and layers.

2.4.4 Future studies

Due to the high density of QTLs in the chicken genome, we did not find more regions containing at least one QTL than would be expected by chance. This may be partially due to the relatively large average size of differentiated regions (1.599Mb), which can be reduced in future studies by use of a denser SNP chip (i.e. the recently developed 600K SNP chip; (Kranis et al., 2013)). A denser SNP chip may also allow the detection of additional regions that were not found in this analysis. Additionally, further studies could use data from whole genome sequencing which may allow the

investigation of non-synonymous changes in the candidate genes discussed above as well as the detection of other selected regions.

2.4.5 Conclusions

In summary, a large number of possible selection signatures were identified by this study of broiler chicken lines. QTLs associated with broiler chicken traits from previous studies have been found in some of these regions. In particular, a region on chromosome 27 was highly differentiated between most of the broiler lines and included a large number of QTL for broiler traits as well as genes associated with growth regulation. Further tests with higher-density marker data may allow these regions to be narrowed down to individual genes.

CHAPTER THREE

Detection of Low Diversity Regions Across Three Broiler Chicken Datasets

3.1 Introduction

In the absence of other factors, positive selection will eventually cause a beneficial allele to become fixed in a population, causing a reduction of diversity in the population at the selected site. Additionally, the hitchhiking effect causes a statistical association between a selected site and neutral sites linked to it, reducing genetic variation around the selected site and resulting in regions of low diversity in the genome (Maynard Smith and Haigh, 1974). Regions of low variation may therefore indicate the action of positive selection. However, regions of low variation may also be created by genetic drift or by demographic events, such as population bottlenecks or founder events.

Such regions of low diversity can be investigated in a number of ways. Several methods calculate the diversity of each SNP in a population and can be used to search for regions of the genome that have zero or very low diversity (Nei, 1987). An alternative method takes advantage of the diversity pattern relative to the position in the genome. As the distance from the selected site increases, the diversity levels in the population should increase. The regression test for this effect fits a regression to the diversity data to test the fit of the data to the theoretical model (Wiener and Pong-Wong, 2011). Other tests take advantage of the linkage disequilibrium created by the hitchhiking effect, such as the long-range haplotype (LRH) test (Sabeti et al., 2002) and the integrated haplotype score (iHS) test (Voight et al., 2006). Both of these tests utilise the extended haplotype homozygosity (EHH) statistic, which involves the probability that two chromosomes that carry the core haplotype are identical by descent (Sabeti et al., 2002). The LRH test is used to find long haplotypes with high

frequency in the population while the iHS test is useful for detecting incomplete sweeps.

Identification of low diversity regions as signatures of selection has been previously applied in livestock using a pooled diversity method. A region of approximately 40kb was found to be nearly fixed for SNPs at the *thyroid stimulating hormone receptor* gene (*TSHR*) in commercial and non-commercial chicken breeds (Rubin et al., 2010). Another study using the same method in a different dataset of chickens found 26 regions showing strong evidence of selection (Elferink et al., 2012). Finally, this pooled heterozygosity method was applied to pig data (Rubin et al., 2012). The most extreme signal was located on chromosome 1, in a region containing major QTL associated with vertebrae number.

Similar regions of low variation have been found in other animals, including a region near the *IGF1* gene in small dogs (Sutter et al., 2007). This gene appears to be a major determinant of small size, with less heterozygosity in this region in small dogs compared to large dogs. A similar pattern was found on a region in chromosome 18 in chondrodysplastic dog breeds relative to other breeds (Parker et al., 2009). Run of homozygosity were recently identified in Holstein cattle and the iHS test was used to identify signatures of selection (Kim et al., 2013). Selection signatures were found on chromosomes 2, 7, 16 and 20, and many of these correspond to the locations of QTL for protein yield, milk and fat content previously found in this breed.

Modern chickens were domesticated before 6000 B.C., but within the last 70 years they have been split into separate lines specialised for egg production (layers) and

meat production (broilers) (Muir et al., 2008). Both types are selected for a wide variety of traits, with layers focused on reproduction and broilers on meat yield and quality. This split was accompanied by large phenotypic changes and the resulting genomic signatures of selection should be detectable by statistical techniques.

In this study, we examined genetic diversity in nine lines of broiler chickens (Andreescu et al., 2007)(Chapter 2) which were genotyped on a 12k SNP chip. Additionally, one of the lines was also genotyped with two higher density SNP arrays, including a 42k panel and a 600k panel. This should allow the detection of regions showing evidence of selection during the recent development of broiler lines.

3.2 Materials and Methods

3.2.1 Animals

In this study, we used birds from nine broiler chicken lines provided by Aviagen, a company primarily involved in broiler breeding. These birds have been selected for a wide variety of traits, including growth rate, body weight, feed conversion ratio, reproduction and welfare traits. Each of these nine lines was subject to slightly different selection criteria, allowing the creation of hybrid lines with different characteristics.

3.2.2 Data

Three sets of data with different SNP densities were used in this study. The first dataset included nine lines of male birds genotyped for 12,046 SNPs, of which 11,988 had known chromosome locations and were distributed across the 28 autosomes and the Z chromosome. This dataset is a subset of the 2.8 million SNPs

identified in the chicken genome sequencing project (Wong et al., 2004), and includes approximately 6000 SNPs genotyped by Andreescu et al. (2007). The distribution of SNPs per chromosome ranged from 26 to 2282 and the mean distance between markers was 0.13 Mb, with a standard deviation of 0.31 Mb (Powell et al., 2012). These markers are well distributed across the genome, but several areas have unusually high or low SNP densities. Chromosome 16 has nearly no coverage, due to the presence of the highly variable major histocompatibility complex. Several low-density areas were also found on chromosomes 1, 2 and 3, while an area on chromosome 3 (110-113Mb) had a particularly high SNP density. This dataset is hereafter referred to as the 12k dataset. As line 3 is present in all datasets, line 3 in the primary focus of this 12k analysis. However, the other 8 lines were also analysed.

The second dataset was genotyped from 161 male and female birds on a chip with a total of 36,454 SNPs. Of these SNPs, 36,366 had known chromosome locations and were distributed across the autosomes and the Z chromosome. The distribution of SNPs per chromosome ranged from 18 to 7164, with chromosome 16 again containing the fewest SNPs. The birds genotyped on this dataset were from line three in the 12k dataset, sampled several years afterwards. The data is hereafter referred to as the 42k dataset.

The final dataset used individuals genotyped from line 3 for 1513 male and female birds on a chip with a total of 580,954 SNPs. Of these SNPs, 553,793 had known chromosome locations and were distributed across the autosomes and the Z chromosome. These SNPs were developed by sequencing 243 individual birds from 24 lines, including broilers, white egg layers, brown egg layers, inbred lines and an unselected layer line. Approximately 139 million SNPs were initially detected, 78

million of which were segregating in one or more lines. This was later reduced to a total of 580,954 SNPs (Kranis et al., 2013). The distribution of SNPs per chromosome ranged from 603 to 102502, with chromosome 16 again containing the fewest SNPs. The birds in this dataset were genotyped at least a year after the 42k dataset. The data is hereafter referred to as the 600k dataset.

3.2.3 Quality Control

Quality control protocols removed closely related individuals, including full sibs, half sibs and parents. The first individual family member found in the pedigree file was retained for analysis. Individuals with more than 10% of SNPs with no calls were also removed. SNPs not called in more than 10% of individuals were removed. Finally, SNPs on the Z chromosome were removed from the analysis. In the 12k dataset, this left an average of 68 individuals per line (70 from line 3) and 11,003 SNPs (see section 2.2.2 for more details). In the 42k dataset, this left 30 individual birds and 32,019 SNPs. Finally, in the 600k dataset, this left 264 individual birds and 530,247 SNPs.

SNPs from both the 12k and 42k datasets were positioned on the WASHUC2 build of the chicken genome. The 600k dataset was positioned on Galgal4, the most recent build of the chicken genome. Therefore, regions identified in the analysis of the 12k and 42k datasets were converted to Galgal4 using the *liftover* software (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

3.2.4 Statistical Analysis - Diversity

To investigate areas of low diversity, Nei's unbiased estimator of nucleotide site diversity (Nei, 1987) was calculated for each SNP within a line for each of the three datasets using the following formula.

$$h = \frac{2n(1 - \sum X_i^2)}{2n - 1}$$

where

$$X_i = X_{ii} + \sum_{j=i} X_{ij}/2$$

n = sample size,

and X_{ij} = frequency of A_iA_j in the sample.

3.2.5 Sliding Windows

In order to reduce stochastic effects, the diversity results for all three datasets were averaged into overlapping sliding windows (Weir et al., 2005). A fixed window size method was used, which keeps a constant window size with a varying number of SNPs per window. Any windows with fewer than 2 SNPs were removed from the analysis to prevent individual SNPs from biasing the results (see section 2.2.4. for more details). It is possible that some windows could be identical to one or more of the previous windows, if no SNPs were lost from the window and no new SNPs were added as the window moved along the genome. "Unique windows" refers to windows which were non-identical.

Each dataset used a separate window size best suited for the individual SNP density. Several window sizes were calculated for each dataset. The size chosen for analysis was based on an average of approximately 10 SNPs present per window. For the 12k

dataset, a window size of 840kbp was used, in which the central positions of the windows were spaced 85kbp apart (see section 2.2.4. for more details). This generated a total of 10498 overlapping sliding windows. For the 42k dataset, a window size of 300kbp and an increment of 30kbp were used, which generated a total of 31,066 overlapping sliding windows. Finally, for the 600k dataset, a window size of 20kbp and an increment of 2kbp were used, which generated a total of 453638 overlapping sliding windows.

3.2.6 Regions of low diversity

In order to identify regions under selection, we investigated windows with zero diversity. “Regions” were defined in the same way as the previous chapter (see section 2.2.6. for more details). Briefly, they were defined by manually searching for windows with zero diversity ("zero diversity windows") physically adjacent to each other. A region could include a small gap of up to two windows to account for windows discarded because they contained fewer than two SNPs. Even with this gap, a large overlap can still be present between adjacent windows, due to the markers shared between adjacent windows in an overlapping sliding window approach. To increase the chances of detecting low diversity regions caused by selection, we also required that regions had to be spread over more than one unique window.

3.2.7 Proportion of zero diversity regions found in each type of region

When a fixed window size is used, the number of SNPs in a region can vary dramatically across the genome. Additionally, the number of individual windows included in a region can be very different, which will change the size of the region. Smaller regions with fewer SNPs will be more likely to include low diversity areas,

as these regions will show greater variance and hence more extreme diversity values. Therefore, it is useful to determine how many zero diversity regions have occurred in regions of a certain size and SNP count.

To investigate this, all regions sizes and SNP counts were recorded for each zero diversity (target) region in each dataset and investigated one at a time. All possible regions of the target size across the genome were investigated using the original increment value for each dataset. The regions with SNP count equal to that of the target region were extracted and the total number of zero diversity regions was counted. The proportion of zero diversity regions out of the total number of regions of the target size and SNP count was calculated. In the 12k dataset where regions of reduced diversity might overlap in multiple lines, the line with the highest proportion of zero diversity regions was used. If there were fewer than 100 regions of the particular size and SNP count in a dataset, the selection criteria was relaxed slightly. The same region size was used but a wider range of SNP counts were extracted, including regions with SNP counts one SNP above or below the target SNP count (e.g. if the target region included 10 SNPs, diversity was assessed for all regions of the target size with 9-11 SNPs). This process was repeated for all region sizes and SNP number combinations found in our zero diversity results for each of the datasets.

3.2.8 Regression test and comparison with diversity results

A regression-based analysis was performed on the 600k dataset, in which asymptotic relationships between diversity and distance from a test position were estimated across the genome using the approach of Wiener and Pong-Wong (2011). This

method detects regions with patterns of variation consistent with positive selection: an asymptotic increase in variability (heterozygosity; *y*) with increasing distance (*x*) from a selected locus is modelled as $y = A + B R^x$ (where R is the asymptotic rate of increase; B is the difference between heterozygosity at the test position and the asymptotic level; A is the asymptotic level of heterozygosity). Positive and increasing regressions ($0<R<1$, $B<0$) were considered as being in the direction expected with positive selection. For this implementation, the test position was moved in steps of 50 Kb across each chromosome and all markers within a fixed distance from this position (bracket sizes: 1, 5, and 10 Mb) were considered in the asymptotic regression curve fitting (Gutiérrez-Gil et al., 2014). A –log(p) value was determined for each test position, where p is the significance level associated with the asymptotic regression. The top 1% of -log(p) values for the respective bracket size were identified as potentially selected regions ("regression regions").

All 600k windows with diversity values < 0.005 ("low diversity") were extracted and combined into regions using the same methods described above for zero diversity regions and then evaluated for correspondence (located within 1Mb) with regression regions.

3.2.9 QTLs and Genes within low diversity regions

We investigated the low diversity regions for QTLs and genes that were previously studied using two resources. For previously identified QTLs, we explored the Animal QTL Database (Hu et al., 2013). The database includes a total of 3,919 QTLs in chickens, representing 297 traits from 192 publications. The base pair positions reported in this database are on the Galgal4 build of the chicken genome. As the SNP

positions for both the 12k and 42k datasets were recorded on the WASHUC2 build of the chicken genome, *liftover* (http://genome.ucsc.edu/cgi-bin/hgLiftOver) was again used to convert positions to Galgal4 using default settings.

The chicken QTL database contains QTLs with large spans. Span is a map property defined by the animal QTL database, which represents a genomic area but does not incorporate any statistical confidence. In order to narrow down the possible QTLs for each region, we assigned the midpoint of each QTL's span as the peak position. We searched the database and included any QTLs whose assigned peak position was located within a region. We used the same method for defining a "broiler QTL" as in the previous chapter (see section 2.2.7. for more details): (1) they were assigned as "production" QTLs in the Chicken QTL DB and (2) they were directly related to meat production rather than egg laying. For example, growth, body weight, feed conversion ratio and breast muscle weight were all considered to be broiler QTL. A full list of broiler traits present in the QTL database can be found in Table S1

To investigate the genes in each region, Ensembl Biomart (http://www.ensembl.org/biomart) was used, with the *Gallus gallus* genes (Galgal4) dataset. The numbers of genes present in the regions and their functions (if available) were recorded.

3.2.10 QTL enrichment

To test if certain regions were enriched for broilers QTLs, a permutation approach was applied. Regions were randomly assigned to positions on the chicken genome 100,000 times. The number of broiler QTL peak positions present in these simulated regions were recorded for each iteration. The top 5% of the 100,000 QTL counts

were determined for each region and used as a significance threshold. Any region with a greater number of QTL than its significance threshold was classed as enriched.

3.2.11 Conversion of base pairs to centiMorgans

Base pair map locations for all three datasets were converted into centiMorgans to allow estimation of the amount of recombination occurring in each region. The most recently published centimorgan map was based on the Galgal3 genome build (Elferink et al. 2010). However a draft centimorgan map for Galgal4 has been created using the following method (Burt and Khoo, personal communication, September 5th, 2014). Estimates of centiMorgan locations for each chromosome on Galgal3 were obtained from Elferink et al. (2010). These locations were updated to the Galgal4 genome build using the SNPs with the same marker ID in the *dbSNP* database (Sherry et al., 2001). Some centiMorgan locations did not have a nearby marker. These locations were estimated using the location of the surrounding markers and the average recombination rate of the chromosome.

3.2.12 Comparison of Datasets

3.2.12.1 Shared regions in line 3: 12k, 42k and 600k

To investigate any regions found in multiple datasets, we searched for zero diversity regions for line 3 found in the same area in the 12k, 42k and 600k datasets. Any region found to overlap in two or more datasets was considered to be a shared region. The region size from the highest density dataset available was recorded in the table. As described above, QTLs, genes and QTL enrichment was investigated for these regions.

3.2.12.2 42K and 12k panels diversity comparison

The diversity of the 12k and 42k datasets in the same genomic locations was directly compared. The 600k panel could not be included, as the 42k and 12k panels were genotyped on the WASUC2 build of the chicken genome and no SNP names were available. To prevent bias, SNPs present in both panels were removed from the 42k dataset. The remaining SNPs in both datasets were averaged into sliding windows of size 840Kbp and increment 85Kbp, the same size as used in the 12k dataset. The diversity found in the same window locations of the 12k and 42k datasets were then plotted against each other, and the correlation calculated.

3.2.12.3 Comparison of low diversity regions in 12k, 42k and 600k datasets

When searching between datasets for overlaps, it is possible that many equivalent regions in the other datasets have very low diversity, but not exactly zero. In order to address this issue, we investigated diversity for the 600k dataset within the zero diversity regions identified in either the 12 or 42k panel. The equivalent windows in the 600k panel were extracted and the diversity investigated. The diversity in these equivalent 12k and 42k windows was then compared to the distribution of 600k diversity windows as a whole. The proportion of 600k windows with diversity less than 0.02 that were identified as zero diversity regions in the 12k or 42k windows was calculated. The same process was repeated when comparing the equivalent regions in the 42k dataset to the 12k zero diversity regions.

3.2.12.4 Comparison with 12k population differentiation

Zero diversity regions and regions found by the regression test were investigated for overlap with high differentiation regions found in line 3 from the previous chapter

(see Chapter 2, Tables 3 and 4 for more details). Theory suggests that low diversity should correlate with high F_{ST}. Any region with an overlap was recorded, including the location of the high differentiation region in Mb and the equivalent in centiMorgans.

3.3 Results

3.3.1 12k regions – Single dataset (Line 3)

Seven regions with zero diversity were found in line 3 in the 12k dataset, located on chromosomes 1, 2, 5 and 14 (Table 1). Five regions were converted to the Galgal4 genome build using *liftover* and the locations of the other two regions were estimated using the position of nearby genes. Regions T12K.1, T12K.2 and T12K.5 were the only zero diversity regions of their size and SNP count and thus the proportion of equivalent zero diversity regions was below 1%. For the other four regions, there were 2 to 21 zero diversity regions of their size and SNP count and thus the proportion of equivalent zero diversity regions ranged from 1.17% to 4.59%. The average zero diversity region size is 1.256Mb (standard deviation = 0.396). None of these line 3 regions equalled or exceeded the broiler QTL significance threshold.

Table 1: Zero diversity regions spread over more than one unique window present in the 12k dataset in line 3, including numbers of SNPs, QTLs, genes and the significance threshold for QTLs. The location includes estimates in both Mb and cM. The final three columns are the number of zero diversity regions of this size and SNP count, the total number of regions of this size and SNP count and the proportion of regions of this size and SNP count which have a diversity of zero.

Region No	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNPs	QTLs	Broiler QTLs	Significance threshold	Genes	Zero diversity regions	Total regions	Proportion of zero diversity
T12K.1	1	54.73	55.89	1.16	102.77	105.65	2.88	19	2	2	10	27	1	331	0.30%
T12K.2	1	158.79	160.89	2.10	299.13	299.49	0.36	6	8	4	15	0	1	169	0.59%
T12K.3	2	52.76	54.30	1.54	124.71	125.01	0.3	5	1	0	11	5	2	119	1.68%
T12K.4	2	92.59	93.47	0.88	189.32	189.68	0.36	3	3	2	8	4	11	650	1.69%
T12K.5	2	97.79	98.81	1.02	195	198.19	3.19	6	4	4	11	14	1	363	0.28%
T12K.6	5	32.62	33.71	1.09	84.7	85.18	0.48	4	3	1	9	9	6	515	1.17%
T12K.7	14	11.55	12.55	1.00	36.42	42.01	5.59	3	3	0	9	43	21	458	4.59%

3.3.2 42k regions

A total of 46 regions with zero diversity were found in the 42k dataset (Table 2). Of the 46 regions identified, 44 were converted directly from WASHUC2 to GALGAL4 using *liftover*. The two remaining regions locations were again estimated using the positions of nearby genes. Regions T42K.23, T42K.29, 32 and T42K.42 were the only zero diversity regions present in regions of their size and SNP count. All other region were one of 2 to 83 zero diversity regions of their size and SNP count (thus representing 0.4% to 20.7% of all equivalent regions).

The average zero diversity region size is 0.518Mb (s.d. = 0.493). Region T42K.8 located on chromosome 1 is the only region that equalled its enrichment threshold of 4 broiler QTLs. This includes two body weight QTLs found in two independent studies, an abdominal fat percentage QTL and a breast muscle weight QTL. This region included 25 genes.

Table 2: Zero diversity regions spread over more than one unique window present in the 42k dataset in line 3, including numbers of SNPs, QTLs, genes and the significance threshold for QTLs. The location includes estimates in both Mb and cM. The final three columns are the number of zero diversity regions of this size and SNP count, the total number of regions of this size and SNP count and the proportion of regions of this size and SNP count which have a diversity of zero. If there columns are highlighted bold, then there were less than 100 total regions, and a range of regions with SNP counts up to one either side of the SNP count was used to increase numbers.

Region No	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNPs	QTLs	Broiler QTLs	Significance threshold	Genes	Zero diversity regions	Total regions	Proportion of zero diversity
T42K.1	1	41.41	41.73	0.32	67.29	67.40	0.11	11	0	0	4	4	83	6001	1.38%
T42K.2	1	54.96	55.94	0.99	103.19	105.73	2.54	34	2	2	9	19	31	2499	1.24%
T42K.3	1	66.32	66.71	0.39	131.85	133.32	1.48	8	0	0	4	4	7	471	1.49%
T42K.4	1	84.41	84.83	0.42	178.39	179.41	1.03	13	0	0	4	12	33	4317	0.76%
T42K.5	1	96.62	97.16	0.53	195.43	195.71	0.28	15	2	2	5	1	7	2144	0.33%
T42K.6	1	134.06	134.61	0.55	263.79	265.09	1.31	17	1	0	5	6	22	3656	0.60%
T42K.7	1	150.14	150.68	0.54	288.70	288.79	0.10	13	0	0	6	1	2	748	0.27%
T42K.8	1	184.87	185.20	0.33	350.64	351.31	0.67	10	4	4	4	25	48	4878	0.98%
T42K.9	1	190.04	190.50	0.45	358.33	358.53	0.20	12	3	1	5	0	2	445	0.45%
T42K.10	2	26.93	27.32	0.39	70.55	70.66	0.11	13	0	0	4	1	64	5301	1.21%
T42K.11	2	44.99	45.34	0.36	111.55	112.71	1.15	8	0	0	4	9	12	891	1.35%
T42K.12	2	52.53	52.72	0.19	124.67	124.70	0.04	7	0	0	2	3	3	113	2.65%
T42K.13	2	58.72	59.05	0.33	130.98	131.54	0.56	12	0	0	4	0	62	5476	1.13%
T42K.14	2	59.25	59.64	0.38	131.90	132.56	0.66	9	8	3	4	2	50	3164	1.58%
T42K.15	2	82.26	82.65	0.38	174.54	174.71	0.16	3	0	0	4	**1**	**41**	**198**	**20.70%**
T42K.16	2	92.27	92.60	0.33	189.18	189.32	0.14	10	5	2	4	2	48	4878	0.98%
T42K.17	2	93.01	93.48	0.47	189.49	189.68	0.19	15	3	2	5	2	14	3213	0.44%
T42K.18	2	93.97	94.30	0.33	189.89	190.03	0.14	6	1	1	4	5	6	361	1.66%
T42K.19	2	94.63	95.04	0.41	190.18	190.36	0.19	13	2	2	4	1	25	3172	0.79%
T42K.20	2	147.50	147.91	0.41	280.17	280.55	0.39	16	0	0	4	3	18	3674	0.49%

T42K.21	3	0.02	0.41	0.39	0.02	0.42	0.40	4	0	0	4	11	**37**	**215**	**17.21%**
T42K.22	3	5.91	6.29	0.38	20.92	21.28	0.35	10	0	0	4	0	13	1959	0.66%
T42K.23	3	79.29	79.77	0.48	169.05	169.61	0.56	19	0	0	5	2	1	2405	0.04%
T42K.24	3	93.83	94.18	0.35	182.82	183.02	0.20	9	0	0	4	1	20	1885	1.06%
T42K.25	4	0.00	0.40	0.40	0.00	0.31	0.31	9	2	0	4	12	8	992	0.81%
T42K.26	4	0.58	0.90	0.32	0.45	0.70	0.25	9	0	0	4	2	50	3164	1.58%
T42K.27	4	24.19	24.57	0.39	62.30	62.83	0.53	10	0	0	4	4	37	3197	1.16%
T42K.28	4	56.12	56.51	0.39	123.12	123.54	0.41	11	0	0	4	3	33	3202	1.03%
T42K.29	4	60.95	64.46	3.51	131.58	134.24	2.67	117	18	9	22	35	1	920	0.11%
T42K.30	4	70.92	71.22	0.31	146.36	146.46	0.10	10	0	0	4	1	48	4878	0.98%
T42K.31	4	86.40	86.73	0.33	177.36	177.69	0.32	8	0	0	4	0	23	1772	1.30%
T42K.32	5	32.59	33.78	1.19	84.69	85.23	0.54	11	5	1	10	13	**1**	**80**	**1.25%**
T42K.33	5	46.29	46.67	0.39	105.41	105.60	0.19	11	3	2	4	1	33	3202	1.03%
T42K.34	5	48.71	49.31	0.60	108.54	109.48	0.94	20	4	2	6	3	47	3778	1.24%
T42K.35	5	55.87	56.20	0.33	141.07	142.90	1.83	12	2	2	4	10	62	5476	1.13%
T42K.36	6	9.66	10.16	0.51	33.29	34.19	0.90	14	0	0	5	7	61	5032	1.21%
T42K.37	6	12.43	12.81	0.39	37.38	37.51	0.13	14	0	0	4	2	50	4897	1.02%
T42K.38	7	35.40	36.11	0.71	102.02	103.00	0.98	22	1	1	6	15	7	2637	0.27%
T42K.39	8	0.18	0.60	0.42	0.10	0.33	0.23	4	0	0	4	0	**32**	**192**	**16.67%**
T42K.40	9	0.21	0.61	0.40	0.00	1.23	1.23	8	0	0	4	17	23	1772	1.30%
T42K.41	10	7.72	8.01	0.29	36.26	36.66	0.40	8	0	0	3	2	23	1772	1.30%
T42K.42	11	2.73	3.11	0.37	6.35	6.53	0.19	14	0	0	4	8	1	295	0.34%
T42K.43	13	9.23	9.54	0.31	31.12	33.41	2.29	6	0	0	4	22	6	219	2.74%
T42K.44	14	0.19	0.69	0.50	0.11	0.40	0.29	15	0	0	5	4	14	3213	0.44%
T42K.45	20	5.42	6.74	1.32	20.98	25.90	4.92	46	1	0	11	43	27	1923	1.40%
T42K.46	21	0.33	0.70	0.37	0.47	1.00	0.53	10	0	0	4	17	37	3197	1.16%

3.3.3 600k Regions

A total of seven regions with zero diversity, which were spread over two or more unique windows, were found in the 600k dataset (Table 3). These regions were present on chromosomes 1, 2, 4 and 5. The average zero diversity region size is 0.031Mb (s.d. = 0.009). For three of the seven regions, the proportion of zero diversity regions of their size and SNP count was less than 0.4%, including regions T600K.1 and T600K.7 that were the only zero diversity regions present in regions of equivalent size and SNP count. For the other four regions, the number of zero diversity regions of equivalent size and SNP count ranged from five to 22. These regions are smaller than regions found in the 12k and 42k datasets, due to the higher SNP density and therefore, fewer QTL peak positions were found in all regions. Additionally, the top 5% enrichment threshold for QTL in each region was zero and therefore an enrichment test was not performed.

Details of all identified low diversity 600K regions, including all QTLs and genes within the regions, can be found in Tables S6 and S7. Some selected examples of these regions are described below. This includes the largest region in terms of physical size, the region containing the most QTL peak positions, the most gene rich region and the region for which the proportion of zero diversity regions of equivalent size and SNP count was smallest.

Table 3: Zero diversity regions spread over more than one unique window present in the 600k dataset in line 3, including numbers of SNPs, QTLs, genes and the significance threshold for QTLs. The location includes estimates in both Mb and cM. The final three columns are the number of zero diversity regions of this size and SNP count, the total number of regions of this size and SNP count and the proportion of regions of this size and SNP count which have a diversity of zero.

Region No	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNPs	QTLs	Broiler QTLs	Significance threshold	Genes	Zero diversity regions	Total regions	Proportion of zero diversity
T600K.1	1	65.972	66.00	0.028	130.54	130.68	0.14	11	0	0	0	2	1	31207	0.0032%
T600K.2	1	77.584	77.608	0.024	154.75	154.78	0.03	3	0	0	0	1	19	870	2.18%
T600K.3	1	139.28	139.328	0.048	276.04	276.05	0.01	4	3	2	0	0	9	329	2.74%
T600K.4	2	52.782	52.808	0.026	124.72	124.72	0.00	3	0	0	0	1	18	693	2.60%
T600K.5	2	135.362	135.386	0.024	267.83	267.87	0.04	4	0	0	0	1	5	1440	0.35%
T600K.6	4	85.26	85.298	0.038	175.90	176.02	0.12	3	0	0	0	7	22	352	6.25%
T600K.7	5	48.906	48.936	0.03	108.76	108.79	0.03	7	0	0	0	0	1	2747	0.04%

3.3.3.1 Region T600K.1 (1: 65.97-66.00Mb) – Region with the smallest proportion of zero diversity windows

Of the 31,207 regions of equivalent size and SNP number, this 0.028Mb (0.14cM) region on chromosome 1 was the only one with zero diversity. While other zero diversity regions were also unique, the proportion of zero diversity windows in region 1 is 0.0032%, suggesting that a zero diversity region of this size and SNP count is particularly rare. Additionally, this is the largest region in centiMorgans on a chromosome with a relatively low recombination rate (Elferink et al., 2010). No QTL peak positions were found in this region, however, two genes were found to overlap it (Table S7, Region 1).

3.3.3.2 Region T600K.3 (1: 139.28-139.33Mb) – Largest region of physical size and region with greatest number of QTL peak positions

A 0.048Mb (0.01cM) region on chromosome 1 was the largest region in terms of physical size. No genes were found in this region. This region also contains the most QTL peak positions (3) of all the 600k regions (Table S6, Region 3). Two of these can be classed as broiler QTL and are associated with body weight and breast muscle weight traits. The proportion of regions of equivalent size and SNP count with zero diversity was 2.74%.

3.3.3.3 Region T600K.6 (4: 85.26-85.30Mb) – Region containing the most genes

A 0.038Mb (0.12cM) region found on chromosome 4 contains seven genes (Table S7, Region 6). No QTLs were found with a peak position in this region. A total of 23 QTLs had spans overlapping the region, nine of which could be classed as broiler traits. These include feed conversion ratio, body weight, abdominal fat weight and

thigh muscle weight. The proportion of regions of equivalent size and SNP count with zero diversity was 6.25%.

3.3.4 600k Regression Analysis

Twenty-one regression regions were identified from one or more bracket size, across 14 chromosomes (bracket size = 1 Mb: 16 regions, bracket size = 5 Mb: 4 regions, bracket size = 10 Mb: 2 regions). Fifteen of these regions were located within 1Mb of a low diversity region (diversity < 0.005) (Table 4). Eight of these regions also overlapped with zero diversity regions from the 12k, 42k or 600k datasets or the highly differentiated regions (identified in Chap. 2) (Table 4). These include region TReg.2 (chrom 1) that overlapped 12k and 42k zero diversity regions (Table 4, Figure 1), region TReg.5 (chrom 2) that overlapped 12k and 42k zero diversity regions (Table 4, Figure 2) and region TReg.7 (chrom 5) that overlapped 42k and 600k zero diversity regions (Table 4, Figure 3). Details of all identified regions in the 600k regression analysis, including all QTLs and genes within the regions, can be found in Tables S8 and S9.

Table 4: 600k windows displaying significant asymptotic regression and diversity less than 0.005 over the three bracket sizes (1Mb, 5Mb and 10 Mb).

Region No	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	Bracket Size (Mb)	QTL	Broiler QTL	Genes	Found in
TReg.1	1	4.526	4.806	0.28	7.58	7.80	0.22	5	0	0	0	N/A
TReg.2	1	54.554	55.670	1.116	102.55	105.29	2.75	1	2	2	22	12k, 42k
TReg.3	1	182.492	182.872	0.38	347.10	347.31	0.21	5	0	0	2	N/A
TReg.4	1	184.152	184.174	0.022	348.13	348.19	0.06	5	0	0	2	N/A
TReg.5	2	26.906	27.314	0.408	70.54	70.66	0.12	1	0	0	2	42k
TReg.6	4	58.526	62.148	3.622	127.14	132.53	5.40	10	6	4	40	F_{ST}
TReg.7	5	48.250	49.418	1.168	108.04	109.76	1.73	1	5	3	15	42k, 600k
TReg.8	5	50.926	50.954	0.028	115.44	115.50	0.05	1	0	0	0	N/A
TReg.9	5	55.928	56.136	0.208	141.63	142.65	1.02	1	1	1	5	42k, F_{ST}
TReg.10	7	35.378	36.214	0.836	*101.97*	*103.00*	1.03	1	1	1	17	42k
TReg.11	11	1.464	4.022	2.558	3.36	6.98	3.62	5, 10	4	2	70	42k
TReg.12	15	3.236	3.568	0.332	3.63	4.05	0.41	1, 5	1	1	4	N/A
TReg.13	20	5.424	5.662	0.238	21.00	22.48	1.48	1	0	0	18	42k
TReg.14	25	0.906	1.762	0.856	36.25	56.22	19.97	1	0	0	84	N/A
TReg.15	28	3.778	4.716	0.938	*52.01*	*54.00*	1.99	1	1	1	42	N/A

Figure 1: Diversity values in region T4.2 on chromosome 1, Region displays asymptotic regression and diversity values less than 0.005 between 54.5 and 56Mb.

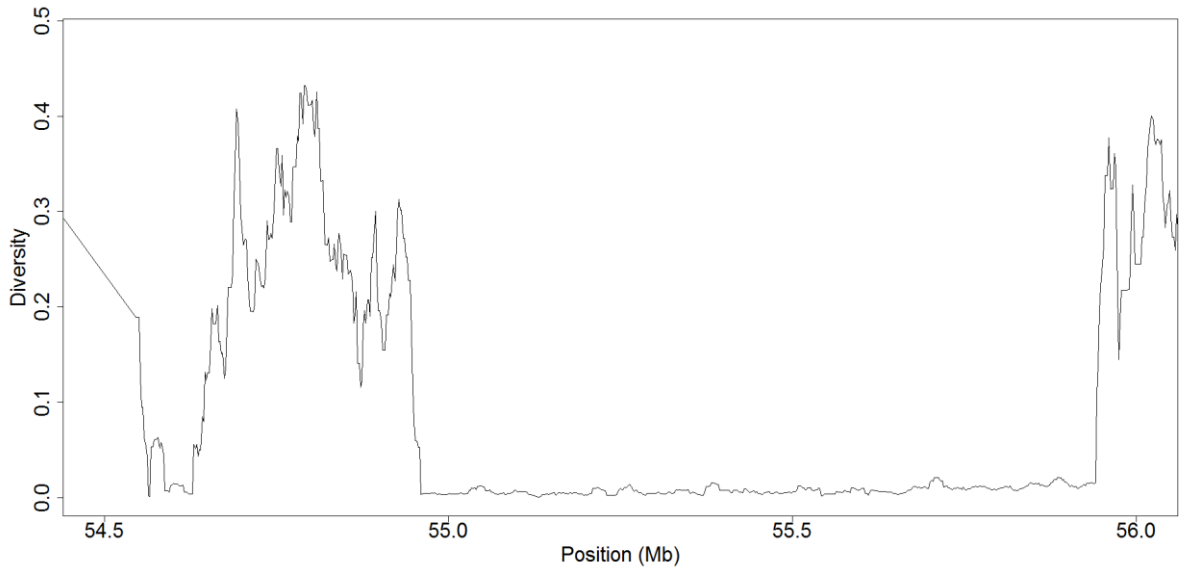


Figure 2: Diversity values in region T4.6 on chromosome 4. Region displays asymptotic regression and diversity values less than 0.005 between 58 and 62Mb.

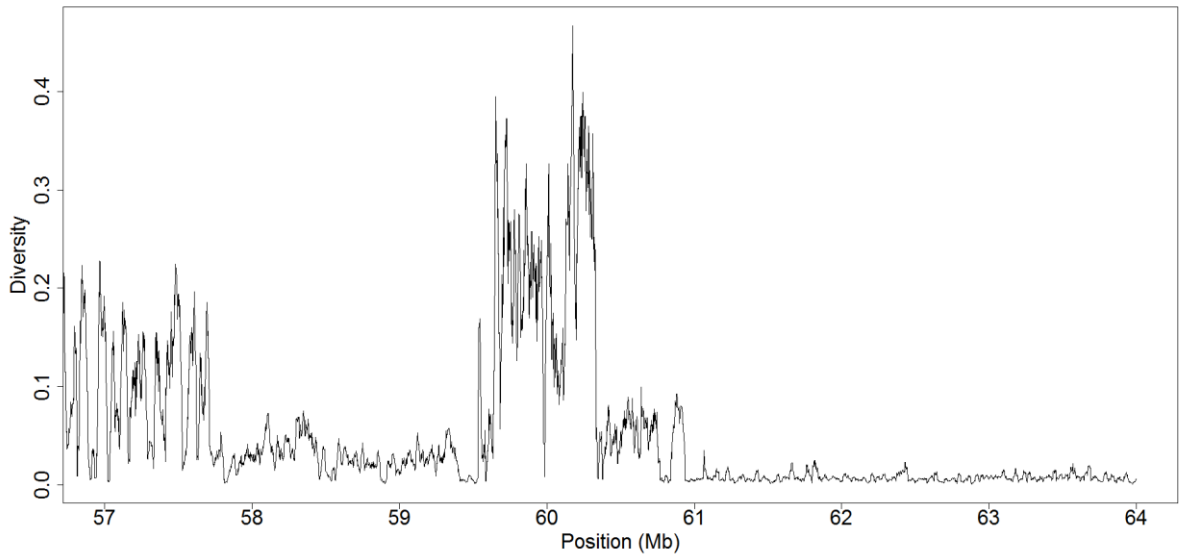
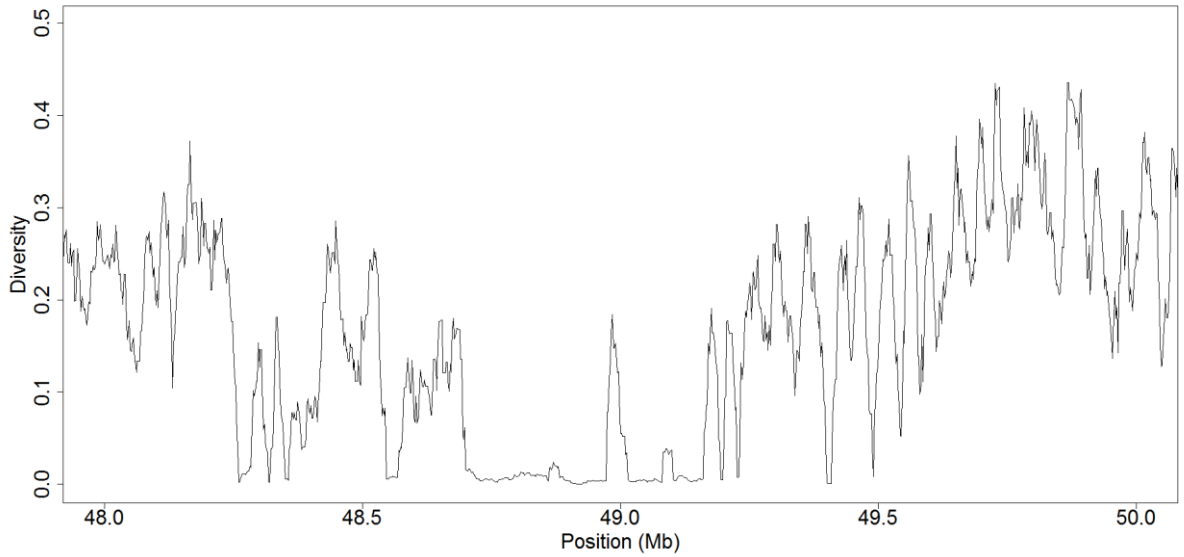


Figure 3: Diversity values in region T4.7 on chromosome 5. Region displays asymptotic regression and diversity values less than 0.005 between 48Mb and 49Mb.



3.3.5 12k regions – All lines

When investigating all nine lines using the 12k dataset, we identified 38 potential regions of selection over thirteen chromosomes (Table 5). Seven of these regions were present in line 3 on chromosomes 1, 2, 5 and 14 (Table 1). For 14 of the 38 regions, the proportion of zero diversity regions of equivalent size and SNP count was less than 1%. Regions T12K9L.2, T12K9L.3, T12K9L.4, T12K9L.8, T12K9L.9, T12K9L.11, T12K9L.14, T12K9L.18, T12K9L.26, T12K9L.34 and T12K9L.38 were the only zero diversity regions present in regions of equivalent size and SNP count. For the other 24 regions, the proportion of zero diversity regions of equivalent size and SNP count ranged from two to 31.

Out of the 38 regions, 23 were found in more than one line. Of these shared regions, lines 1 and 7 had the most signals of selection with 15 regions. Lines 2, 4 and 9 had only two shared regions, despite their close relationship found in the previous chapter (see section 2.3.1. for more details). Lines 3, 6 and 8 appear together much more frequently (seven regions). Line 5 appears in six regions. The average region size is

1.21Mb (s.d. = 0.35). Region T12K9L.15 is a low diversity region on chromosome 2 identified in all nine lines and may represent selection which occurred before the separation of the lines. One haugh unit QTL (a reproduction trait related to the correlation of the height and weight of the albumen in the egg) was found in this region.

Table 5: Zero diversity regions spread over more than one unique window present in the 12k dataset in all nine lines, including numbers of SNPs, QTLs, genes and the significance threshold for QTLs. Regions present in line 3 are highlighted (bold).

Region No.	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	Lines	SNPs	QTLs	Broiler QTLs	Significance threshold	Genes	Zero diversity regions	Total regions	Proportion of zero diversity
T12K9L.1	1	4.06	4.90	0.83	7.21	7.88	0.67	1,2,7	3	0	0	8	6	29	650	4.46%
T12K9L.2	1	53.26	54.26	1.00	97.92	102.17	4.25	9	5	1	1	9	18	1	539	0.19%
T12K9L.3	1	53.85	55.39	1.54	101.03	104.65	3.62	1	12	5	5	11	27	1	417	0.24%
T12K9L.**4**	**1**	**54.73**	**55.89**	**1.17**	102.77	105.65	2.88	**3, 8**	19	**2**	**2**	**10**	**27**	1	331	0.30%
T12K9L.5	1	66.32	67.85	1.53	131.86	137.36	5.5	6	4	2	0	11	24	9	289	3.11%
T12K9L.6	1	70.07	71.80	1.74	142.74	148.29	5.55	8	4	26	16	12	25	8	157	5.10%
T12K9L.7	1	73.67	74.58	0.91	149.42	149.66	0.24	2	3	1	1	8	10	8	175	4.57%
T12K9L.8	1	114.78	116.92	2.15	227.88	229.69	1.81	1,5,7	6	6	3	16	9	1	179	0.56%
T12K9L.**9**	**1**	**158.79**	**160.89**	**2.10**	299.13	299.49	0.36	**1,3,5,7,8,9**	6	**8**	**4**	**15**	**0**	1	169	0.59%
T12K9L.10	1	161.87	163.19	1.33	299.66	299.88	0.22	1	4	13	9	11	3	18	379	4.75%
T12K9L.11	1	186.54	187.53	1.00	353.94	355.62	1.68	1,7	18	2	2	9	12	1	309	0.32%
T12K9L.12	2	25.91	26.91	1.00	70.26	70.54	0.29	1,7	4	2	0	9	6	9	573	1.57%
T12K9L.13	2	26.67	27.66	1.00	70.47	70.76	0.28	6	3	1	0	9	6	23	581	3.96%
T12K9L.14	2	28.60	30.19	1.59	71.11	72.91	1.8	1,7	10	0	0	12	16	1	412	0.24%
T12K9L.**15**	**2**	**52.76**	**54.30**	**1.54**	124.71	125.01	0.3	**1,2,3,4,5,6,7,8,9**	5	**1**	**0**	**11**	**5**	2	119	1.68%
T12K9L.16	2	56.34	57.75	1.41	127.61	130.21	2.6	2,4,9	3	0	0	11	8	12	281	4.27%
T12K9L.**17**	**2**	**92.59**	**93.47**	**0.88**	189.32	189.68	0.36	**3,8**	3	**3**	**2**	**8**	**4**	16	650	2.46%
T12K9L.**18**	**2**	**97.79**	**98.81**	**1.02**	195	198.19	3.19	**1,3,6,7**	6	**4**	**4**	**11**	**14**	1	363	0.28%
T12K9L.19	3	81.74	82.65	0.91	172.51	173.59	1.07	1,7	3	2	0	8	9	25	650	3.85%
T12K9L.20	4	24.73	26.66	1.93	63.01	63.57	0.55	1,5,8	5	0	0	14	7	5	158	3.16%
T12K9L.21	4	25.58	27.08	1.50	63.25	63.69	0.43	2	4	0	0	11	1	7	289	2.42%
T12K9L.22	4	27.51	28.76	1.25	63.81	65.59	1.78	1,5,6	5	0	0	10	7	2	387	0.52%

T12K9L.23	4	31.05	32.22	1.18	72.27	74.92	2.65	7	3	5	2	10	8	25	458	5.46%
T12K9L.24	5	2.52	3.96	1.45	10.3	10.82	0.52	6	4	0	0	11	21	10	330	3.03%
T12K9L.**25**	**5**	**32.62**	**33.71**	**1.09**	84.7	85.18	0.48	**3,4,8**	4	**3**	**1**	**9**	**9**	6	515	1.17%
T12K9L.26	6	13.12	14.62	1.50	37.61	38.39	0.78	1	7	3	0	11	16	1	368	0.27%
T12K9L.27	6	23.60	24.94	1.34	57.8	59.54	1.74	1,5,7	3	0	0	11	11	23	325	7.08%
T12K9L.28	6	29.25	30.34	1.09	71.91	76.07	4.16	8	3	1	0	9	20	20	509	3.93%
T12K9L.29	7	4.64	5.55	0.91	17.59	22.79	5.2	5	3	4	4	8	19	13	650	2.00%
T12K9L.30	8	24.16	25.40	1.24	82.11	87.26	5.15	2,6,8	3	1	0	10	16	20	391	5.12%
T12K9L.31	10	6.60	7.46	0.86	35.01	35.92	0.91	1,7	3	0	0	8	18	31	581	5.34%
T12K9L.32	11	1.80	2.65	0.85	5.32	6.31	0.99	1,7	5	2	0	8	37	3	500	0.60%
T12K9L.33	11	6.36	7.17	0.81	13.23	15.81	2.57	1,7	5	0	0	7	10	6	619	0.97%
T12K9L.34	11	7.26	8.57	1.31	16.04	23.68	7.64	1,7	14	4	1	11	17	1	424	0.24%
T12K9L.35	12	11.25	12.19	0.94	41.28	45.89	4.61	1	6	10	6	9	44	7	655	1.07%
T12K9L.36	14	1.02	1.94	0.92	0.6	1.62	1.03	1,7	4	1	0	9	14	8	617	1.30%
T12K9L.**37**	**14**	**11.55**	**12.55**	**1.00**	36.42	42.01	5.59	**2,3,5,8**	3	**3**	**0**	**9**	**43**	21	458	4.59%
T12K9L.38	20	6.59	7.50	0.91	25.73	26.88	1.16	8	8	2	2	8	23	1	548	0.18%

3.3.6 Comparisons of datasets

3.3.6.1 Line 3: Regions identified by multiple SNP panels

A total of five regions were shared between at least two of the three datasets (Table 6) Regions TShared.1, TShared.3 and TShared.4 were found on chromosomes 1, 2 and 5 in the 12k (line 3) and 42k datasets. Region TShared.5 was found on chromosome 5 in the 42k and 600k datasets. Finally, region TShared.2 was found on chromosome 2 in the 12k (line 3) and 600k datasets. Three regions contained some broiler QTL peak positions. Regions TShared.2 and TShared.5 did not contain any QTL peak positions but this may be due to their small size. None of these regions exceeded their enrichment threshold for numbers of QTL.

Table 6: Zero diversity regions spread over more than one unique window shared between at least two of the three datasets, including numbers of QTLs, genes and the significance threshold for QTLs. The start and end points displayed here are from the population differentiation table.

Shared Region No.	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNPs	QTLs	Broiler QTLs	Threshold	Genes	Datasets
TShared.1	1	54.96	55.94	0.99	103.19	105.73	2.54	34	2	2	9	19	12k, 42k
TShared.2	2	52.782	52.808	0.026	124.72	124.72	0.00	3	0	0	0	1	12k, 600k
TShared.3	2	92.27	92.60	0.33	189.18	189.32	0.14	10	5	2	4	2	12k, 42k
TShared.4	5	32.59	33.78	1.19	84.69	85.23	0.54	11	5	1	10	13	12k, 42k
TShared.5	5	48.906	48.936	0.03	108.76	108.79	0.03	7	0	0	0	0	42k, 600k

3.3.6.2 Comparison of zero diversity regions to highly differentiated regions

In chapter 2, a number of highly differentiated regions were identified using Weir and Cockerham's F_{ST} estimator (Weir and Cockerham, 1984) and a circular permutation method (Cabrera et al., 2012). Five highly differentiated regions were found to overlap with regions of zero diversity in the 42k zero diversity and 600k regression results (Table 7). Region TF_{ST}.4 on chromosome 5 overlapped with both 42k region T42K.35 and regression region TReg.9 and is therefore an interesting candidate region. Region TF_{ST}.2 and TF_{ST}.3 overlapped with regression region TReg.6 on chromosome 4. There were no overlaps between the highly differentiated regions and 12k (line 3) or 600k zero diversity regions.

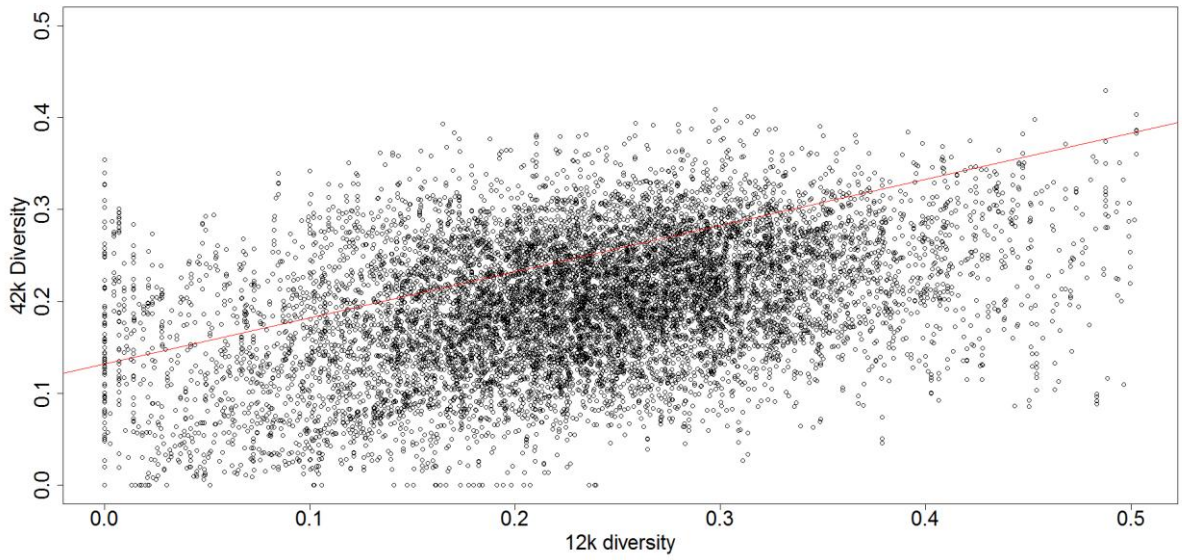
Table 7: Zero diversity regions spread over more than one unique window in all datasets also found in the high differentiation results in the previous chapter (Chapter 2). There were no overlaps with the 12k and 600k zero diversity regions.

Region Number	F_{ST} Region number	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	Overlaps with
TF_{ST}.1	Shared Region 5	1	133.97	136.13	2.16	263.69	268.90	5.21	42k Region 6
TF_{ST}.2	Shared Region 24	4	57.62	59.21	1.59	125.88	127.96	2.08	600k Regression R6
TF_{ST}.3	Shared Region 25	4	58.64	59.69	1.06	127.27	128.99	1.72	600k Regression R6
TF_{ST}.4	Shared Region 29	5	55.37	56.90	1.53	137.37	144.91	7.55	42k Region 35, 600k Regression R9
TF_{ST}.5	Shared Region 47	20	6.17	7.34	1.17	25.25	26.67	1.43	42k Region 45

3.3.6.3 Comparison of diversity in 12k and 42k SNP panels

When the two sets of diversity results at each window were plotted against each other (Figure 4) we see that the 12k dataset captures a large proportion of the diversity seen in the higher density panel. The correlation between the two datasets was 0.395. The fitted line on Figure 4 reveals that there are several of zero diversity regions in the 12k (line 3) dataset which contain more diversity in the 42k dataset, due to the higher available SNP density.

Figure 4: Comparison of diversity in the same window locations in the 12k and 42k datasets. SNPs found in both the 42k and 12k dataset were removed from the 42k before comparison to prevent bias. The remaining SNPs in both datasets were averaged into sliding windows of size 840Kbp and increment 85Kbp. Windows in the same genomic location were plotted against each other for comparison. The 12k dataset captures a large proportion of the available diversity, despite the relatively low SNP density. The datasets show a correlation of 0.395.



3.3.6.4 Distribution of diversity

There were only seven regions with zero diversity in the 600k dataset, and only one of these was also present in the 42k dataset. However, there were many more individuals present in the 600k dataset, and thus a higher probability that at least one

individual bird in the population would display some diversity at a SNP within a sliding window by chance. This may be due to a mutation in an individual bird or the allele not reaching complete fixation. To further investigate the correspondence between the two datasets, the 600k windows in the corresponding locations to the 42k zero diversity regions were investigated. Approximately 38% of 600K windows with diversity values less than 0.02 were equivalent to the zero diversity regions found in the 42k dataset (Figure 5a). Diversity in the 600k dataset also investigated in the 12k (line 3) zero diversity regions (Figure 5b). However, the proportion of windows in corresponding locations here was much smaller (5.32%), possibly because there are only seven 12k line 3 zero diversity regions available for comparison. The 42k windows equivalent to the 12k (line 3) zero diversity regions were compared in the same way (Figure 5c). Of 42K windows with diversity values less than 0.02, 7.33% were equivalent to the zero diversity regions found in the 12k dataset

Figure 5a: Comparison of overall 600k diversity (blue) and 600k diversity in regions equivalent to the 42k regions (purple).

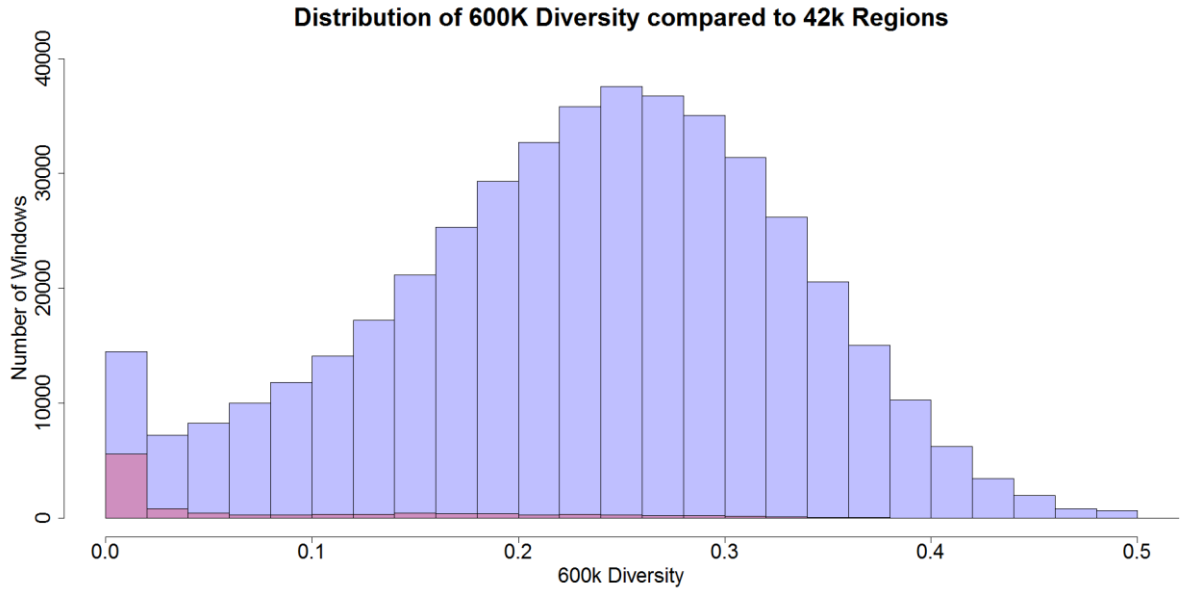


Figure 5b: Comparison of overall 600k diversity (blue) and 600k diversity in regions equivalent to the 12k regions in line 3 (purple).

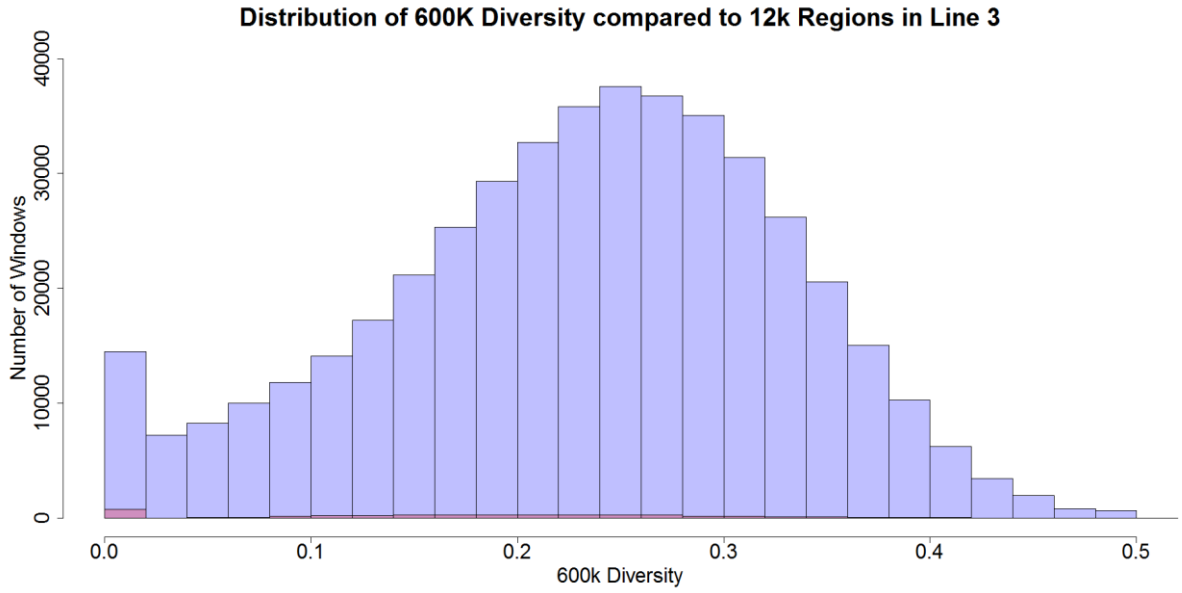
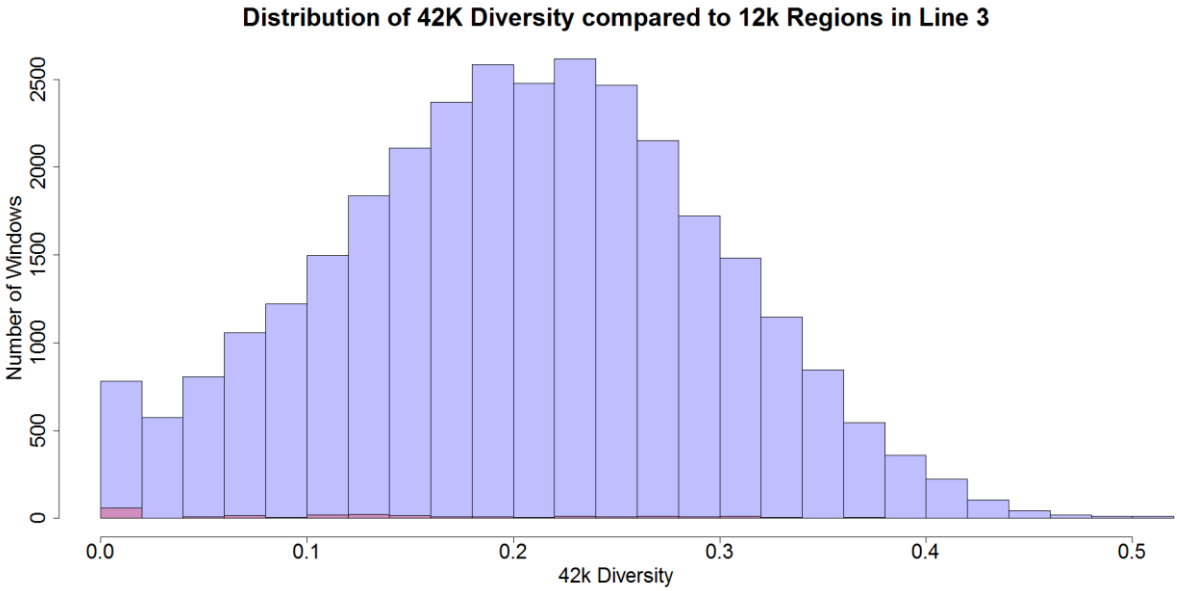


Figure 5c: Comparison of overall 42k diversity (blue) and 42k diversity in regions equivalent to the 12k regions in line 3 (purple).



3.4.1 Discussion

Domesticated chickens, specialised for their role in either egg or meat production, were produced via artificial selection. This selection caused large phenotypic changes and the resulting genomic signatures of selection are detectable by statistical techniques. In this study, we have identified putative signatures of selection in broiler chickens by analysing genetic diversity patterns in three datasets of genome wide SNPs (12k, 42k and 600k). All three SNP panels were examined in a single broiler line (line 3) while eight additional lines were analysed for the 12k panel. A total of seven low diversity regions were found in the 12k (line 3) dataset, 46 were found in the 42k dataset and seven were found in the 600k dataset. Three regions were identified for all three panels in line 3. Fifteen candidate regions were found in the 600k dataset using the regression method, which fits a regression to the diversity data. This tests the fit of the data to the theoretical model of selective sweeps (as the distance from the selected site increases, the diversity levels in the population should increase). Eight of these regression regions overlapped with zero diversity regions in other datasets, indicating they are strong candidates for selection. Additional zero diversity regions were found when all nine lines in the 12k dataset were included in the analysis. This includes one region found in all nine lines, which may indicate older selection which occurred before these nine separate lines were created.

3.4.2 Sliding Window method

When using a sliding window method, usually a fixed number of SNPs is used as the size of the window. The windows move along the genome one SNP at a time (see section 2.2.4. for more details). As the 12k dataset contained extremely variable SNP

densities in different genomic areas, this method would have resulted in considerable variation in window size. It would therefore have been difficult to accurately compare windows across the genome. Instead a fixed window size method was used, which allows the SNP density to vary across the genome while the windows remain at the same size. In order to prevent windows with very low SNP density from affecting the results, windows with fewer than 2 SNPs were removed from the analysis. Regions were determined by searching for consecutive unique windows of zero diversity, allowing for small gaps of up to two windows.

This fixed window method and the method of determining regions means that the identified regions can vary in both size and SNP count. Certain size and SNP count combinations are more likely to include regions of zero diversity. Small regions with only 3 SNPs will be more likely to have a diversity of zero than large regions with 10 SNPs. To deal with this issue we calculated how many regions of the same size and SNP count had a diversity of zero. Several regions were the only zero diversity regions with their size and SNP count. This includes regions T600K.1 and T600K.7 of the 600k zero diversity dataset. T600K.1 was the only zero diversity region among 31,207 regions with the same size and SNP count. This also applies to regions T12K.1, T12K.2 and T12K.5 in the 12k line 3 dataset and regions T42K.23, T42K.29, T42K.32 and T42K.42 in the 42k dataset.

3.4.3 Composition of Regions

In the 12k (line 3) dataset, five of the seven regions contain peak positions of QTL relating to broiler characteristics found in previous studies. In the 42k dataset, 15 of the 46 regions contain peak positions of broiler QTL. While these are relatively low

numbers, many more broiler QTL are present if the entire QTL span is taken into account. Of the five regions shared between at least two datasets, three (TShared.1, TShared.3 and TShared.4) contain broiler QTL.

3.4.4 Candidate selected regions

Due to the greater resolution offered by the 600k dataset and the use of the regression based analysis, we focussed on candidate regions in the 600k zero diversity regions and 600k regression regions. Region TReg.2 on chromosome 1 contains the gene *insulin-like growth factor 1 (IGF1),* which has been shown to be associated with growth, skeletal integrity and feeding traits in chickens (Zhou et al., 2005, Amills et al., 2003). *IGF1* was also identified as a region of low diversity using a pooled heterozygosity method in brown egg layer lines (Qanbari et al., 2012). This region also overlapped with zero diversity regions in the 12k and 42k datasets (TShared.1) Region TReg.10 on chromosome 7 contains the gene *glycerol-3-phosphate dehydrogenase 2* (*GPD2*). This gene is associated with increased glycerol production, leading to an accumulation of fat and was found to be expressed at higher levels in lean chickens in two lines divergently selected for abdominal fat content (Resnyk et al., 2013). This region also overlapped with a zero diversity region in the 42k dataset.

Of the 15 regression regions, eight were also found to overlap with zero diversity regions for one of the SNP panels and three of these were found to overlap with zero diversity regions from at least two panels. A number of these regions overlapped with regions showing evidence of selection in one or more previous studies of chickens (Rubin et al., 2010, Elferink et al., 2012, Zhang et al., 2012a, Zhang et al.,

2012b). Region TReg.11 overlapped with zero diversity regions found in the 42k dataset. Additionally, it overlapped with regions of high homozygosity in a set of domestic lines and a subset of commercial broilers (Rubin et al., 2010) and a set of 67 commercial breeds (Elferink et al., 2012). There were also overlaps with regions found in two lines divergently selected for abdominal fat content (Zhang et al., 2012a). This region is very gene rich, overlapping 70 genes and containing the peak position for two broiler QTL (Table S9, Region 11), which are related to abdominal fat content. However, none of these genes were obvious functional candidates for broiler traits. Additionally, this 2.5Mb region spans approximately 3.62cM so that recombination may be relatively common in this region.

Another regression region, region TReg.6 on chromosome 4, is also a candidate region. This region overlaps with a highly differentiated region found in the previous chapter as well as regions of high homozygosity found in commercial broilers (Elferink et al., 2012, Rubin et al., 2010, Qanbari et al., 2012). Forty genes overlap with this region (Table S9, Region 6) as well as four broiler QTL related to feed intake, shank length and meat quality (Table S8, Region 6). One of these genes is *melatonin receptor 1A*. Melatonin has been associated with reproductive traits in chickens, including skeletal development in the eggs of laying hens (Taylor et al., 2013), number of eggs and age at first egg (Li et al., 2013).

Regression region TReg.13 found on chromosome 20 overlaps with a zero diversity region in the 42k dataset, as well as a local reduction in heterozygosity in pooled sequence data from laying chickens (Qanbari et al., 2012). 18 genes were found in this region (Table S9, Region 13), but none were functional candidates for broiler traits. However, a *growth hormone releasing hormone precursor* gene was found

within 1Mb of the region. Mutations in growth hormone genes are associated with small size in mice (Godfrey et al., 1993). No broiler QTL peak positions were found in this region.

Region T600K.1 is the only zero diversity region of its size and SNP count of a total of 31,207 regions. Additionally, it has the largest size in centiMorgans of all the 600k zero diversity regions, despite being located on chromosome 1, a chromosome with relatively low recombination. This makes this region a good candidate for a putative selected sweep. However, there are no broiler QTLs located in this region. There are two micro RNAs (gga-mir-6608-1 and gga-mir-6608-2) which overlap with this region but their function is unknown (Table S7, Region 1). 26 genes were found within 1Mb of the region but none were functional candidates for broiler traits.

Region T12K9L.15 was found in all nine lines in the 12k dataset and may represent selection which occurred before the lines separated. Five genes were found in this region which includes *TPK1*, *7SK* and 3 uncharacterised proteins. These genes were not considered to be functional candidates for broiler traits. A QTL relating to haugh units was found in this region, which is associated with the height and weight of the albumen in the egg (Baker et al., 1996). This suggests this region may be related to reproduction traits, which are selected in broilers as well as layers.

3.4.5 Comparison of diversity across the SNP panels

In the 600k dataset, only seven regions of zero diversity were found. When investigating the regions shared between datasets, only one region was found on the 600k dataset. There are several reasons for this small number of discovered regions. The 600k dataset contains a much larger number of animals than the smaller datasets

and a much larger number of SNPs. Therefore, there is a much greater chance that at least one of the birds in the 600k dataset will contain some variation. As a strict criterion of zero diversity was used, regions where only one bird in the population contained a small amount of variation would not be identified.

In order to investigate this, diversity in the 600k regions which correspond to the 42k dataset zero diversity regions was analysed. 1499 windows were found in these regions with a diversity score under 0.005, which is approximately 38% of the total regions under 0.005 diversity. This suggests many more 600k regions would overlap with regions from the 42k dataset if the zero diversity criteria were relaxed to very low diversity. There were fewer regions when comparing the 600k and 12k line 3 regions in the same way. This may be due to the small number of 12k (line 3) regions found. Few regions were also found when comparing the 42k and 12k line 3 regions. The diversity in the 12k and 42k datasets was also compared directly at each window location. We find a correlation of 0.395 even when SNPs common to both datasets have been removed. This suggests that the 12k dataset is capturing a lot of the diversity of the 42k dataset, despite its lower resolution.

Five regions were found that overlapped between two of the three zero diversity datasets. No zero diversity region was found in all three datasets. There were a large number of zero diversity regions in the 42k dataset. This is probably due to the small number of individual birds present in that dataset after quality control (30) compared to the 12k line 3 (68) and the 600k datasets (264). Because of this, it is useful to understand how rare the individual zero diversity regions are in the 42k dataset. There were a total of five regions which were in the 42k dataset which were the only zero diversity regions found of their size and SNP count.

3.4.6 SNP selection

All three of these SNP datasets were designed to be used in commercial chicken breeds and therefore were segregating in one or more lines (Wong et al., 2004, Andreescu et al., 2007, Kranis et al., 2013). This suggests that the zero diversity regions found here are more likely to be due to more recent, line-specific selection rather than older, domestication based sweeps. An exception to this is region T12KL9.15, which occurs in all nine lines and is therefore could be due to older selection which occurred before the lines were separated.

3.4.7 Future Studies

In order to narrow down the regions to the strongest candidates for broiler selection, only regions with a diversity of zero were considered. However, this will have prevented us from detecting sweeps where the selected allele is not fixed, e.g. due to recent selection. Also, additional criteria were applied to the regions, so that the region had to occur in at least two consecutive unique windows. This may hide smaller regions of selection. Relaxing these criteria in the future may allow the identification of more regions. Finally, using data from whole genome sequencing could allow the identification of more selected regions. Additionally, this data would allow the investigation of non-synonymous SNP changes within the candidate regions identified described above.

3.4.8 Conclusions

A number of putative selection signatures were identified in nine broiler chicken lines. In some of the regions, QTLs related to broiler traits were found by previous studies. Several regions were found in both the regression analysis and at least one of

the zero diversity dataset analyses. One of these regions contains the gene *IGF1*, known to be involved in growth and a likely candidate gene for broiler selection. Several other regions were found which overlapped with selection signatures identified by previous studies. More signatures of selection may be found by including regions of low diversity instead of only zero diversity regions.

CHAPTER FOUR

Detecting Signatures of Selection between Broiler and Layer Chickens

4.1 Introduction

Modern chickens (*Gallus gallus domesticus*) are primarily bred for either egg production (layers) or meat production (broilers). Both lines are selected for a wide variety of traits. Selection on layers has focussed on reproduction traits, such as age at time of first egg-laying, while broilers have been selected for meat production traits, including body weight, abdominal fat content and feed conversion ratio. This split into two separate groups occurred within the last 70 years (Muir et al., 2008), and has caused large phenotypic changes whose effects on patterns of genetic variability should be detectable by statistical techniques.

QTL mapping has been used previously to identify genomic regions associated with traits of biological interest. To identify QTL In chickens associated with meat or egg production, QTL mapping was carried out using broiler and layer crosses. Four main QTL regions associated with broiler traits were found on chromosome 1 in an F_2 chicken line produced by crossing broiler and layer lines (Nones et al., 2006). The same F_2 population was later used to identify QTL related to carcass quality, including carcass percentage, breast percentage and wings percentage (Baron et al., 2011). Another F_2 population was produced from the Ross 308 broiler line and a White Leghorn layer line, and used to locate several body weight QTL (Sewalem et al., 2002). The same population was later used to identify QTL for bone mineral traits (Podisi et al., 2012), age and body weight at first egg (Podisi et al., 2011), feather growth and body weight (Hocking et al, 2012) as well as candidate genes related to follicle number (McDerment et al, 2012)

However, QTL mapping requires extensive phenotypic information and rarely provides a precise localisation of the variants responsible for phenotypic differences (Haley, 1995). An alternative method is to use genotypic information to detect genomic regions displaying signatures of selection, which may be associated with traits of interest. There are a number of widely used methods for detecting signatures of selection, including population differentiation (Akey et al., 2002, Lewontin and Krakauer, 1973) and local reduction in diversity (Rubin et al., 2010, Qanbari et al., 2012).

Population differentiation involves differences in allele frequencies between populations and can be quantified by the F_{ST} statistic, as discussed in Chapter 2. In order to determine which loci are affected by selection, an outlier method can be used (Akey et al., 2002). The method identifies selected regions as the outliers in the upper tails of the empirical distribution of F_{ST} and removes the need for neutral simulations. This technique has been used in a number of species, including dogs (Akey et al., 2010), cattle (Flori et al., 2009), sheep (Arora et al., 2011) and pigs (Wilkinson et al., 2013). Estimates of the levels of population differentiation between broiler and layer populations may provide information concerning selective signatures created after the separation of these lines.

Local reductions in diversity can also signal the effects of positive selection, since this can cause alleles to rise to high frequencies or fixation within a population. The hitchhiking effect of a positively selected allele (a selective sweep) can create a statistical association between the neutral sites linked to the selected site, causing a low diversity region around the selected site (Maynard Smith and Haigh, 1974). This technique will capture both selection which increases the frequency of an allele

which has been segregating in a population (soft sweep) and selection which increases the frequency of a new mutation (hard sweep) (for more details, see section 1.2). However, in lines with a relatively small population size, both types of sweep will appear to be hard sweeps, even if the selected allele was previously segregating in the population (Jensen, 2014). Several studies that have identified low diversity regions have been carried out in chickens; for example the *TSHR* gene has been identified as a low diversity region using a pooled heterozygosity method in several studies (Rubin et al., 2010, Qanbari et al., 2012) (see section 1.3.2 for more details).

In this study, we examined genetic differentiation between a broiler and layer line using genome wide data from a 42k SNP chip (described in Chapters 2 and 3) to identify regions where selection may have taken place during broiler and layer specialisation. A selective sweep in a genomic location within one of the lines can result in increased population differentiation, as sweeps reduce local diversity and hence can increase F_{ST} in the region surrounding the selected variant (Charlesworth, 1998, Cruickshank and Hahn, 2014). As in Chapter 3, we therefore also examined genetic diversity across the genome of the layer line used in the study. Genomic regions that were candidates for signatures of selective sweeps were identified by both approaches.

4.2 Materials and Methods

4.2.1 Animals

In this study, we used birds from two commercial chicken lines provided by Aviagen and Hy-Line. Aviagen is a company primarily involved in broiler breeding while Hy-Line is involved in layer breeding. One of these lines consists of broilers, which are

specialised for meat production and are selected for a number of traits, including growth rate, feed conversion ratio, body weight and meat quality, as well as some reproduction and welfare based traits. This broiler line is the same as line 3, which was used in the previous chapters. The second line consists of brown egg layers, which are specialised for egg production and are also selected for a number of traits, including age at first egg, egg number and egg shell thickness as well as other welfare based traits

4.2.2 Data

Both lines were genotyped on the same SNP chip. The broiler line was genotyped from 161 male and female birds while the brown egg layer line was genotyped from a total of 365 female birds. This SNP chip consists of a total of 36,454 SNPs, 36,366 of which had a known chromosome location distributed across the autosomes and the Z chromosome. The number of SNPs per chromosome ranged from 18 to 7164. Chromosome 16 contains the fewest SNPs, due to the presence of the highly variable major histocompatibility complex (Delany et al., 2009). Chromosome 1 contains the most SNPs, as it is the largest macrochromosome in the genome.

4.2.3 Quality Control

In previous chapters quality control protocols removed closely related individuals and individuals with more than 10% of SNPs with no calls. There was no pedigree information available for the layer line so closely related individuals were not removed in this study. In both the broiler and layer datasets, all individual birds had at least 90% of their SNPs present, so all were retained for analysis. SNPs with no calls in more than 10% of individuals were removed from the analysis. Additionally,

SNPs located on the Z chromosome were removed. This left a total of 31449 SNPs for analysis.

4.2.4 Statistical Analysis

4.2.4.1 Population Differentiation

To investigate areas of high population differentiation between the layer and broiler datasets, Weir and Cockerham's F_{ST} (Akey et al., 2002, Weir and Cockerham, 1984) was calculated for each SNP using the following formula.

$$F_{ST} = \frac{MSP - MSG}{MSP + (n_c - 1)MSG}$$

where MSG denotes the observed mean square for loci within *s* populations and MSP denotes the observed mean square between populations:

.

$$MSG = \frac{1}{\sum_{i=1}^{s} n_i - 1} \sum_{i}^{s} n_i P_{A\mathrm{i}} (1 - P_{Ai})$$

$$MSP = \frac{1}{s - 1} \sum_{i}^{s} n_i (P_{A\mathrm{i}-\bar{P}_A})^2$$

n_c is the average sample size across samples that accounts for the variance in sample size over subpopulations.

$$n_c = \frac{1}{s - 1} \sum_{i=1}^{s} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$$

n_i denotes the sample size in subpopulation i.

P_{Ai} denotes the frequency of SNP allele A in the ith subpopulation.

$\bar{p}$ is the weighted average of P_A across subpopulations.

$$\bar{p} = \frac{\sum n_i P_{Ai}}{\sum n_i}$$

4.2.4.2 Diversity

To investigate areas of low diversity in the layer dataset, Nei's unbiased estimator of nucleotide site diversity (Nei, 1987) was calculated for each SNP in the layer dataset using the following formula. Low diversity regions in the broiler dataset were investigated in chapter 3 (See section 3.3.2. for more details).

$$h = \frac{2n(1 - \sum X_i^2)}{2n - 1}$$

where

$$X_i = X_{ii} + \sum_{j=i} X_{ij}/2$$

n = sample size, and X_{ij} = frequency of A_iA_j in the sample.

4.2.5 Sliding Windows

Both the population differentiation and diversity results were averaged into overlapping sliding windows to reduce stochastic effects. A fixed window size method was used, so each window was always the same physical size but the number of SNPs within a window could vary. If a window contained two SNPs or fewer, it

was removed from the analysis. This was to prevent windows with one SNP from biasing the results. The windows moved along the genome in a fixed increment size. This means that it was possible that some windows could contain the same exact SNPs as the previous window. "Unique windows" refers to windows that were non-identical. A window size of 300kbp was used, in which the central position of the windows were spaced 30kbp apart. This is the same window size and increment used for the 42k dataset in the previous chapter (See section 3.2.5 for more details) and is based on an average of approximately 10 SNPs present per window.

4.2.6 Regions with high differentiation between broilers and layers

To detect regions with high differentiation, we applied the outlier method to the F_{ST} values, as used in Chapter 2. This method identified the upper tail of the empirical distribution of the F_{ST} windows, which was defined as the top 0.5% of F_{ST} window values. This is equivalent to F_{ST} values greater than 0.7026. Regions were defined the same way as previous chapters. Regions consist of physically adjacent windows where the F_{ST} values exceed the top 0.5% threshold. There could be a small gap of up to two windows, to account for window removed for containing fewer than 2 SNPs. Even with this gap, a large overlap is still present between the windows due to shared markers in the overlapping sliding windows. To further reduce the likelihood that a high F_{ST} region was detected by random chance, a region had to contain more than one unique window.

4.2.7 Regions with low diversity – layer dataset

To identify putative selected regions in the layer dataset, we investigated windows with zero diversity. Such regions were defined the same way as the high

differentiation regions, using physically adjacent windows of zero diversity. They could contain a gap of two windows and had to contain more than one unique window.

4.2.8 Proportion of highly differentiated or zero diversity regions found in each type of region

The fixed size sliding window method allows the number of SNPs in an individual window to vary. Additionally, the method of defining regions allows any number of windows (above two) to be included in a region. This means the size and SNP counts in different regions can be very different. Smaller regions containing few SNPs are more likely to contain low diversity or highly differentiated values, as these regions will show greater variance and hence more extreme values.

It is therefore useful to determine how many highly differentiated or zero diversity regions have occurred in regions of a certain size and SNP count. All region sizes and SNP counts were recorded for each putative (target) region in both the population differentiation and diversity regions. All possible regions of the target size across the genome were investigated using the original increment of 30kbp. The possible regions with a SNP count equal to the target SNP count were extracted and the total number of highly differentiated regions (region with F_{ST} values greater than 0.7026) or zero diversity regions were counted. The proportion of zero diversity or high differentiation regions out of the total number of regions of the target size and SNP count was calculated. If there were 100 regions or fewer of the target size and SNP count, the criterion was slightly relaxed. A wider range of SNP counts were used with the same region size, including regions with SNP counts one SNP above

and below the target SNP count (e.g. if the target region included 10 SNPs, diversity or high differentiation was assessed for all regions of the target size with 9-11 SNPs). This was repeated for all region sizes and SNP count combinations found in both the high differentiation and zero diversity results.

4.2.9 QTLs and Genes within putative selection regions

Highly differentiated and low diversity regions were investigated for QTLs and genes found in previous studies using two resources. The Animal QTL Database (http://www.animalgenome.org/cgi-bin/QTLdb/index) was used to explore previously identified QTLs (Hu et al., 2013). In chickens, this database contains a total of 3,919 QTLs, representing 297 traits from 192 publications, recorded on the Galgal4 genome build, the most recent version of the chicken genome assembly (International Chicken Genome Sequencing Consortium, 2004). As both datasets were originally recorded on the WASHUC2 (Galgal3) chicken genome build, they were converted to Galgal4 using *liftover* (http://genome.ucsc.edu/cgi-bin/hgLiftOver) with default settings (for more details, see sections 2.2.7 and 3.2.3).

Rather than using the whole "span" for each QTL recorded in the chicken genome database, the midpoint of the span was assigned as a "peak position". QTLs with peak positions found within a region could be defined as broiler or layer QTL. Broiler QTLs were assigned as "production" QTLs in the Chicken QTL DB and were directly related to meat production. This includes traits such as breast muscle weight, abdominal fat percentage, body weight and feed conversion ratio. A full list of broiler traits present in the QTL database can be found in Table S1. Layer QTLs were also assigned as production traits, and were directly related to egg production.

This includes traits such as egg shell thickness, age at first egg and number of eggs. A full list of layer traits present in the QTL database can be found in Table S10.

Ensembl Biomart (http://www.ensembl.org/biomart) was used to identify genes overlapping with each region, using the *Gallus gallus* genes (Galgal4) dataset. The numbers of genes present in the regions and their functions (if available) were recorded. Each gene overlapping a region was investigated as a candidate for broiler or layer traits.

4.2.10 QTL Enrichment

A permutation approach was applied to test whether any regions were enriched for QTL. Each region was assigned to random positions on the chicken genome 100,000 times. The number of QTL peak positions present in these permuted regions was recorded for each iteration. This was treated as a null distribution for each region, and the top 5% of this distribution was used as an enrichment threshold. Any region that equalled or exceeded its enrichment threshold was classed an enriched (similar to the approach used in sections 2.2.8. and 3.2.10).

4.2.11 Conversion of base pairs to centiMorgans

To estimate the amount of recombination occurring in each region, base pair locations were converted into centiMorgans using the linkage map described in the previous chapter (See section 3.2.11 for more details). This linkage map was produced using a Galgal3 linkage map (Elferink et al., 2010) converted to Galgal4 by locating the equivalent marker positions in the *dbSNP* database (Sherry et al., 2001). Some centiMorgan locations did not have a nearby marker and were estimated using

the location of the surrounding markers and the average recombination rate of the chromosome (Burt and Khoo, personal communication, September 5th, 2014).

4.2.12 Comparisons with previous results

Both the highly differentiated and low diversity layer regions were compared to results from the previous two chapters to find any overlaps between regions, including highly differentiated regions in nine lines of broiler chickens (Chapter 2), zero diversity regions in nine lines of broiler chickens genotyped on 12k SNPs, zero diversity regions in one broiler line genotyped for 42k and 600k SNPs and regions displaying significant asymptotic regressions and low diversity in the broiler line genotyped for 600k SNPs (Chapter 3).

4.3 Results

4.3.1 Highly Differentiated Regions

A total of 32 regions with high levels of differentiation between the 42k broiler and layer datasets were found (Table 1). They were located on chromosomes 1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13 and 20. Twenty-six regions were converted to the Galgal4 format using *liftover*. The locations of the other six regions were estimated using the positions of nearby genes. The average length of a highly differentiated region is 0.37Mb (s.d = 0.068). Regions F_{ST}.10, F_{ST}.13, F_{ST}.30 and F_{ST}.32 were the only highly differentiated regions found in regions of their specific sizes and SNP counts. For the other regions, the number of highly differentiated regions of equivalent size and SNP count range from two to 19 and the proportion of highly differentiated regions of equivalent size and SNP count ranged from 0.07% to 9.66%. No regions were found to be significantly enriched for QTL.

Table 1: High differentiation regions in the top 0.5% of F_{ST} windows and spread over more than one unique window present in the broiler and layer datasets, including numbers of SNPs, QTLs and genes. The location includes estimates in both Mb and cM. The average diversity in the region in both the broiler and layer lines is included for comparison. The final three columns are the number of highly differentiated equivalent regions of this size and SNP count, the total number of regions of this size and SNP count and the proportion of regions of this size and SNP count which are highly differentiated.

Region No.	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNP count	Broiler Diversity	Layer Diversity	QTLs	Broiler QTLs	Layer QTLs	Genes	High Differentiation Regions	Total Regions	Proportion with High differentiation
F_{ST}.1	1	0.47	0.80	0.33	0.58	0.98	0.40	11	0.20	0.10	0	0	0	16	15	6013	0.25%
F_{ST}.2	2	27.14	27.50	0.35	70.61	70.71	0.10	10	0.04	0.16	1	0	1	1	15	3524	0.43%
F_{ST}.3	2	42.43	42.79	0.36	102.03	102.95	0.92	11	0.18	0.12	1	1	0	2	11	4990	0.22%
F_{ST}.4	2	69.80	70.11	0.31	167.10	167.15	0.05	7	0.10	0.22	0	0	0	0	13	980	1.33%
F_{ST}.5	2	81.34	81.67	0.33	174.16	174.29	0.14	10	0.05	0.20	1	0	0	0	11	5168	0.21%
F_{ST}.6	3	17.69	18.05	0.36	60.26	62.06	1.80	8	0.10	0.18	0	0	0	5	10	1069	0.94%
F_{ST}.7	3	51.28	51.63	0.36	125.84	127.26	1.41	11	0.16	0.07	1	0	0	12	11	4990	0.22%
F_{ST}.8	3	69.31	69.64	0.32	152.52	152.65	0.13	7	0.16	0.16	0	0	0	1	13	980	1.33%
F_{ST}.9	3	88.16	88.51	0.36	178.36	178.57	0.21	12	0.08	0.23	0	0	0	0	10	5649	0.18%
F_{ST}.10	3	93.77	94.30	0.53	182.79	183.14	0.35	15	0.04	0.12	0	0	0	1	1	2391	0.04%
F_{ST}.11	4	56.90	57.23	0.33	123.95	124.66	0.70	12	0.08	0.25	1	1	0	0	10	5200	0.19%
F_{ST}.12	5	5.95	6.31	0.36	16.41	17.72	1.32	5	0.23	0.10	0	0	0	4	2	127	1.57%
F_{ST}.13	5	27.64	28.15	0.51	79.54	81.14	1.60	17	0.16	0.17	1	1	0	6	1	4304	0.02%
F_{ST}.14	5	31.79	32.32	0.53	84.40	84.59	0.20	3	0.22	0.11	3	0	0	5	14	145	9.66%
F_{ST}.15	5	46.56	46.88	0.33	105.54	105.70	0.16	11	0.03	0.19	2	1	0	0	15	6013	0.25%
F_{ST}.16	5	49.92	50.25	0.33	112.85	114.06	1.21	16	0.13	0.15	0	0	0	9	2	494	0.40%
F_{ST}.17	5	50.52	50.84	0.33	114.62	115.28	0.66	14	0.07	0.22	1	0	1	4	2	1788	0.11%
F_{ST}.18	6	20.17	20.47	0.29	44.80	45.06	0.26	12	0.04	0.26	0	0	0	6	11	5649	0.19%
F_{ST}.19	8	9.83	10.25	0.42	39.66	39.80	0.15	6	0.21	0.14	1	1	0	8	3	104	2.88%
F_{ST}.20	10	1.84	2.17	0.33	12.63	20.29	7.66	5	0.06	0.10	0	0	0	38	4	224	1.79%
F_{ST}.21	10	3.69	4.07	0.38	29.92	30.90	0.98	13	0.29	0.08	0	0	0	4	8	5304	0.15%
F_{ST}.22	10	4.40	4.87	0.47	31.72	32.44	0.72	16	0.17	0.03	0	0	0	6	3	4585	0.07%
F_{ST}.23	10	9.91	10.18	0.28	39.93	40.41	0.48	8	0.24	0.00	0	0	0	10	10	1069	0.94%
F_{ST}.24	10	13.73	14.06	0.33	56.44	57.82	1.38	9	0.20	0.14	0	0	0	3	19	3534	0.54%
F_{ST}.25	10	15.47	15.76	0.29	63.26	63.81	0.56	11	0.10	0.19	0	0	0	0	15	6013	0.25%
F_{ST}.26	11	1.82	2.23	0.41	5.36	6.10	0.74	14	0.07	0.19	1	0	0	22	6	5070	0.12%
F_{ST}.27	11	17.51	17.87	0.36	57.13	59.28	2.15	9	0.13	0.15	2	0	0	7	7	2048	0.34%
F_{ST}.28	12	15.13	15.47	0.34	54.71	55.62	0.91	10	0.08	0.16	0	0	0	9	2	750	0.27%
F_{ST}.29	13	0.54	0.87	0.33	0.54	0.87	0.33	7	0.09	0.20	0	0	0	3	13	980	1.33%

F_{ST}.30	13	9.06	9.45	0.39	30.52	32.52	2.00	7	0.06	0.01	4	2	2	11	1	299	0.33%
F_{ST}.31	13	10.95	11.40	0.45	38.37	39.16	0.79	11	0.18	0.13	0	0	0f	6	2	1271	0.16%
F_{ST}.32	20	6.00	6.47	0.47	25.05	25.59	0.54	15	0.00	0.14	1	0	0	4	1	4242	0.02%

Details of all identified regions, including all QTLs and genes within the regions, can be found in Tables S11 and S12. Some selected examples of these regions are described below. This includes the largest region in terms of physical size, the largest region in terms of centiMorgans, the region containing the most QTL peak positions, the most gene rich region, and the region for which the proportion of high differentiation regions of equivalent size and SNP count was smallest.

4.3.1.1 Region F_{ST}.10 (3: 93.77 - 94.30Mb) - Largest region of physical size

A 0.53Mb (0.35cM) region on chromosome 3 was the largest region in terms of physical size (Figure 1). The *SOX11* gene was found in this region which is associated with skeletal development (Table S12, Region 10). No QTL peak positions were found. Among the 2391 regions of equivalent size and SNP count, this was the only highly differentiated region and thus the proportion of highly differentiated regions of equivalent size and SNP count to this region was 0.04%.

Figure 1: F_{ST} and diversity in region F_{ST}.10 located on chromosome 3.

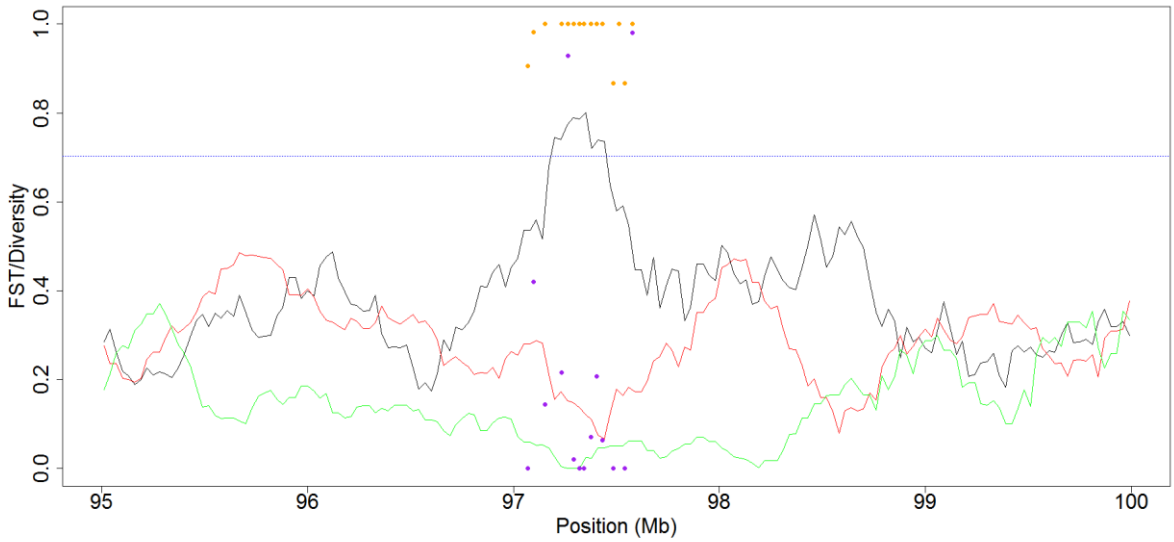


Figure 1 displays the F_{ST} window values calculated between the broiler and layer datasets (black line), the diversity in the layer dataset (red line), the diversity in the broiler dataset (green line), the top 0.5% of F_{ST} values threshold (blue dashed line) and the allele frequencies of individual SNPs within the selected regions in the layer dataset (purple dots) and broiler dataset (orange dots) between approximately 95 and 100Mb on chromosome 3. Mb locations are on the Galgal3 genome build. To allow easier visual comparison, the displayed allele frequency is the dominant allele in the broiler dataset.

4.3.1.2 Region F_{ST}.20 (10: 1.84 - 2.17Mb) - largest region in terms of centiMorgans and region containing the most genes

A 0.33Mb region on chromosome 10 is the largest region in terms of centiMorgans (7.66cM) (Figure 2). This suggests that the recombination rate in this region is well above the average recombination rate of 3.4cM/Mb for this chromosome (Elferink et al, 2010). No QTL peak positions were found in this region. However, this is the most gene rich of the highly differentiated regions, with 38 overlapping genes (Table

S12, Region 20). Of the 224 equivalent regions of this size and SNP count, four were highly differentiated, thus the proportion of highly differentiated regions of equivalent size and SNP count was 1.79%.

Figure 2: F_{ST} and diversity in regions F_{ST}.20, F_{ST}.21, F_{ST}.22 and Div.3 located on chromosome 10.

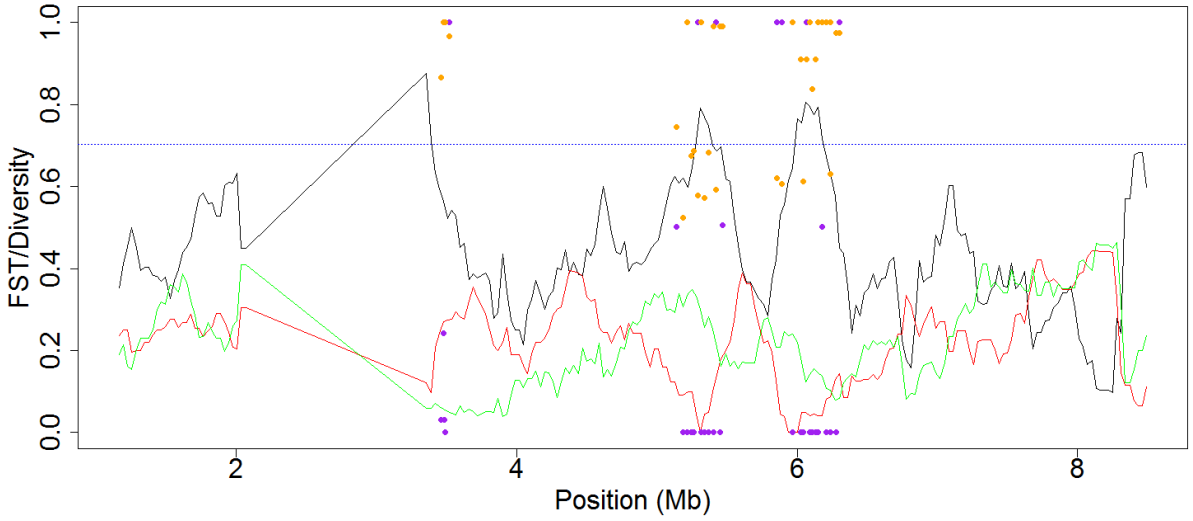


Figure 2 displays the F_{ST} window values calculated between the broiler and layer datasets (black line), the diversity in the layer dataset (red line), the diversity in the broiler dataset (green line), the top 0.5% of F_{ST} values threshold (blue dashed line) and the allele frequencies of individual SNPs within the selected regions in the layer dataset (purple dots) and broiler dataset (orange dots) between approximately 1.5 and 8.5Mb on chromosome 10. Mb locations are on the Galgal3 genome build. To allow easier visual comparison, the displayed allele frequency is the dominant allele in the broiler dataset.

4.3.1.3 Region F_{ST}.30 (13: 9.06 - 9.45Mb) - region with greatest number of QTL peak positions

A 0.39Mb (2cM) region on chromosome 13 is the region with the greatest number of QTL peak positions (Figure 3). This region contains four QTLs. Two QTLs are for

age at first egg and are classed as layer traits (Table S11, Region 30). The other two QTLs are for body weight and drumstick weight, and are classed as broiler traits. Eleven genes were found in this region. Of the 299 equivalent regions of this size and SNP count, this was the only highly differentiated region and thus the proportion of highly differentiated regions of equivalent size and SNP count was 0.33%.

Figure 3: F_{ST} and diversity in region F_{ST}.30 and F_{ST}.31 located on chromosome 13.

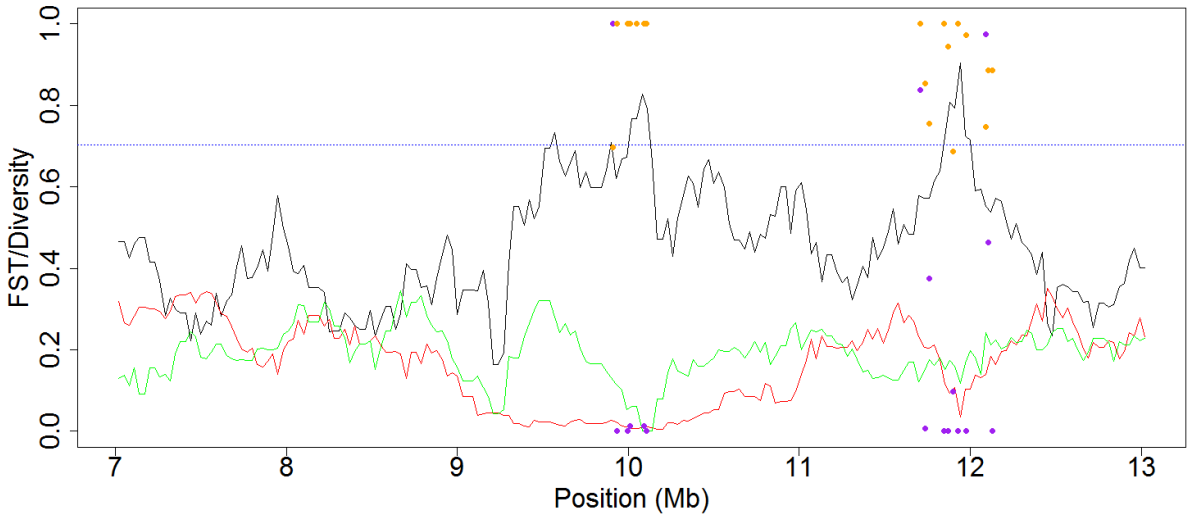


Figure 3 displays the F_{ST} window values calculated between the broiler and layer datasets (black line), the diversity in the layer dataset (red line), the diversity in the broiler dataset (green line), the top 0.5% of F_{ST} values threshold (blue dashed line) and the allele frequencies of individual SNPs within the selected regions in the layer dataset (purple dots) and broiler dataset (orange dots) between approximately 7 and 13Mb on chromosome 13. Mb locations are on the Galgal3 genome build. To allow easier visual comparison, the displayed allele frequency is the dominant allele in the broiler dataset.

4.3.1.4 Region F_{ST}.13 (5: 27.64 - 28.15Mb) - Region with the smallest proportion of high differentiation regions

A 0.51Mb (1.6cM) region on chromosome 5 had the smallest proportion of high differentiation regions (Figure 4). Of a total of 4304 equivalent regions of this size and SNP count, this was the only highly differentiated region and thus the proportion of highly differentiated regions of equivalent size and SNP count was 0.02%. Six genes were found in this region (Table S12, Region 13). One broiler QTL peak position was found in this region, which is related to residual feed intake.

Figure 4: F_{ST} and diversity in region F_{ST}.13 located on chromosome 5.

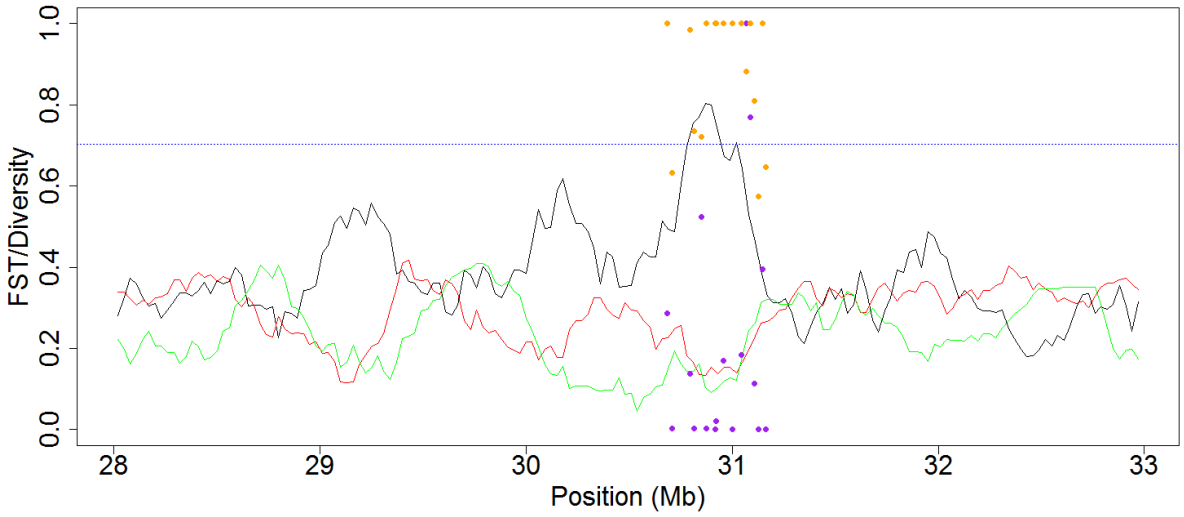


Figure 4 displays the F_{ST} window values calculated between the broiler and layer datasets (black line), the diversity in the layer dataset (red line), the diversity in the broiler dataset (green line), the top 0.5% of F_{ST} values threshold (blue dashed line) and the allele frequencies of individual SNPs within the selected regions in the layer dataset (purple dots) and broiler dataset (orange dots) between approximately 28 and 33Mb on chromosome 5. Mb locations are on the Galgal3 genome build. To

allow easier visual comparison, the displayed allele frequency is the dominant allele in the broiler dataset.

4.3.2 Zero Diversity Layer regions

A total of seven zero diversity regions that were spread over two or more unique windows were found in the layer dataset (Table 2). These regions were present on chromosomes 3, 5, 10 and 11. The average zero diversity region is 0.38Mb (s.d. = 0.08). Regions Div.4 (Figure 5), Div.5 (Figure 5), and Div.7 were the only zero diversity regions present in regions of equivalent size and SNP count. For the other regions, the numbers of highly differentiated regions of equivalent size and SNP count range from two to four and the proportion ranged from 0.04% to 1.52%. No regions were found to be enriched for QTL.

Table 2: Zero diversity regions spread over more than one unique window present in the layer dataset, including numbers of SNPs, QTLs, genes and the significance threshold for QTLs. The location includes estimates in both Mb and cM. The final three columns are the number of zero diversity regions of this size and SNP count, the total number of equivalent regions of this size and SNP count and the proportion of regions of this size and SNP count which have a diversity of zero.

Region No.	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNP count	QTLs	Broiler QTLs	Layer QTLs	Enrichment threshold	Genes	Zero Diversity Regions	Total Regions	Proportion with Zero Diversity
Div.1	3	49.73	50.05	0.33	123.31	123.65	0.34	9	0	0	0	8	3	3	3534	0.08%
Div.2	5	5.92	6.25	0.33	16.31	17.47	1.15	3	0	0	0	8	2	4	263	1.52%
Div.3	10	4.34	4.69	0.35	31.58	32.22	0.65	11	0	0	0	9	5	2	4990	0.04%
Div.4	10	8.76	9.23	0.48	38.40	39.03	0.63	18	1	0	0	14	3	1	4133	0.02%
Div.5	10	9.91	10.41	0.51	39.93	40.83	0.91	14	0	0	0	14	21	1	274	0.36%
Div.6	11	12.21	12.53	0.32	33.09	33.37	0.28	11	0	0	0	8	3	2	4990	0.04%
Div.7	11	12.63	12.94	0.31	33.46	33.73	0.27	9	2	2	0	7	2	1	2048	0.05%

Details of all identified regions, including all QTLs and genes within the regions, can be found in Tables S13 and S14. Some selected examples of these regions are described below, chosen in the same way as above.

4.3.2.1 Region Div.5 (10: 9.91 - 10.41Mb) - Largest region of physical size and region containing the most genes

A 0.51Mb (0.91cM) region on chromosome 10 was both the largest region of physical size and the most gene rich (Figure 5). The region contains 21 genes (Table S14, Region 5) but no QTL peak positions. Of the 263 equivalent regions of this size and SNP count, there were four zero diversity windows, thus the proportion zero diversity regions of equivalent size and SNP count to this region was 1.52%.

Figure 5: F_{ST} and diversity in regions Div.4, F_{ST}.23 and Div.5 located on chromosome 10.

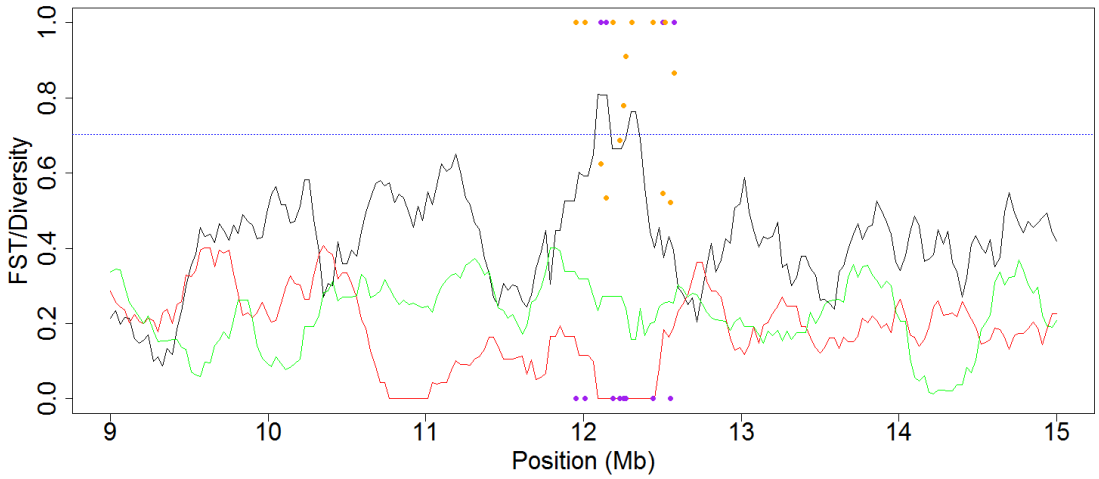


Figure 5 displays the F_{ST} window values calculated between the broiler and layer datasets (black line), the diversity in the layer dataset (red line), the diversity in the broiler dataset (green line), the top 0.5% of F_{ST} values threshold (blue dashed line) and the allele frequencies of individual SNPs within the selected regions in the layer dataset (purple dots) and broiler dataset (orange dots) between

approximately 9 and 15Mb on chromosome 10. Mb locations are on the Galgal3 genome build. To allow easier visual comparison, the displayed allele frequency is the dominant allele in the broiler dataset.

4.3.2.2 Region Div.2 (5: 5.92 - 6.25Mb) - largest region in terms of centiMorgans

A 0.33 Mb (1.15cM) region on chromosome 10 was the largest region in terms of centiMorgans (Figure 6). The region contains two genes (Table S14, Region 2) but no QTL peak positions. Of the 274 equivalent regions of this size and SNP count, this was the only zero diversity region, thus the proportion of zero diversity regions of equivalent size and SNP count was 0.36%.

Figure 6: F_{ST} and diversity in regions F_{ST}.12 and Div.2 located on chromosome 5.

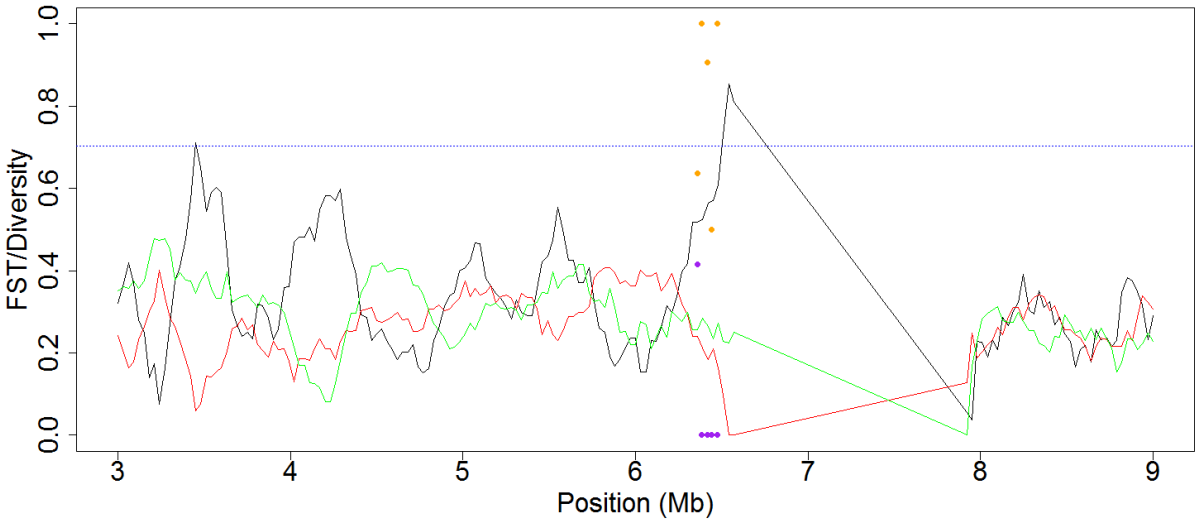


Figure 6 displays the F_{ST} window values calculated between the broiler and layer datasets (black line), the diversity in the layer dataset (red line), the diversity in the broiler dataset (green line), the top 0.5% of F_{ST} values threshold (blue dashed line) and the allele frequencies of individual SNPs within the selected regions in the layer dataset (purple dots) and broiler dataset (orange dots) between

approximately 3 and 9Mb on chromosome 5. Mb locations are on the Galgal3 genome build. To allow easier visual comparison, the displayed allele frequency is the dominant allele in the broiler dataset.

4.3.2.3 Region Div.7 (11: 12.63 - 12.94Mb) - region with greatest number of QTL peak positions

A 0.31Mb (0.27cM) region located on chromosome 11 contained two QTL peak positions, associated with body weight and growth, from two separate studies (Table S13, Region 7). These traits are associated with broilers rather than layer chickens. Of the 2048 equivalent regions of this size and SNP count, this was the only zero diversity region. The proportion zero diversity regions of equivalent size and SNP count was 0.05%.

4.3.2.4 Region Div.10 (10: 8.76 - 9.23Mb) - Region with the smallest proportion of zero diversity regions

A 0.48Mb (0.63cM) region on chromosome 10 was the region with the smallest proportion of zero diversity regions. Of a total of 4133 equivalent regions of this size and SNP count, this was the only zero diversity region, thus the proportion of zero diversity regions of equivalent size and SNP count was 0.02%. Three genes were found in this region, as well as one QTL (Tables S13 and S14). However, it was associated with heart size, which is not an obvious production trait.

4.3.3 Comparison of regions

4.3.3.1 Overlaps between the population differentiation and zero diversity regions

Three regions located on chromosomes 5 and 10 were found to overlap between the broiler and layer population differentiation regions and the zero diversity layer

regions (Table 3). The smallest region size in Mb was recorded in Table 3. These regions are fixed for one allele in layers and are fixed (or nearly fixed) for the alternative allele in broilers. The allele frequencies in region F_{ST}.23/Div.5 (Figure 5) show that of the 14 SNPs in the region, four are fixed for the alternative allele in broilers and another SNP is close to fixation (allele frequency > 90%). Only three SNPs in this region have a similar allele frequency in both broilers and layers. Of the four SNPs in region F_{ST}.12/Div.2 (Figure 6), two are fixed for the alternative allele in broilers and layers and another is nearly fixed. Finally, of the 11 SNPs in region F_{ST}.22 (Figure 2), three were fixed for the alternative allele in broilers and another three were nearly fixed. Only two of the 11 SNPs had similar allele frequencies in broilers and layers. A similar pattern can be found in other high differentiation regions, including F_{ST}.10 (Figure 1), F_{ST}.13 (Figure 4), F_{ST}.21 (Figure 2), $F_{ST.}$30 and $F_{ST.}$31 (Figure 3). This may indicate that selection has occurred in these regions in both broilers and layers.

Table 3: Regions overlapping between the broiler and layer F_{ST} results (Table 1) and the zero diversity layer results (Table 2). The smaller region location of the F_{ST} and zero diversity results is recorded in the table. This table includes the location in both Mb and cM, the number of SNPs in the region and the number of QTL and genes present.

Diversity Region	F_{ST} Region	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNP count	QTLs	Broiler QTLs	Layer QTLs	Enrichment threshold	Genes
Div.2	F_{ST}.12	5	5.92	6.25	0.33	16.31	17.47	1.15	3	0	0	0	8	2
Div.3	F_{ST}.22	10	4.34	4.69	0.35	31.58	32.22	0.65	11	0	0	0	9	5
Div.5	F_{ST}.23	10	9.91	10.18	0.28	39.93	40.41	0.48	14	0	0	0	6	10

4.3.3.2 Overlaps of the selected regions with zero diversity, high differentiation and significant asymptotic regression regions identified in broilers

In the high differentiation regions between the broiler and layer datasets, 13 of the 32 regions overlapped with zero diversity, high differentiation and significant asymptotic regressions regions identified for the broiler lines analysed in previous chapters (Table 4)(see chapter 2, tables 3 and 4 and chapter 3, tables 1, 2, 3, 4 and 5, for more details). Seven of these overlaps were with the 12k high F_{ST} regions between nine broiler lines, four overlaps were with the 12k (broiler) zero diversity regions, five overlaps were with the 42k (broiler) zero diversity regions and two overlaps were found with the 600k regression regions. The most prominent of these was region F_{ST}.2 on chromosome 2, which overlapped with the 12k and 42k (broiler) zero diversity and 600k regression regions. For the zero diversity layer analysis, two of the seven identified regions overlapped with the 12k F_{ST} regions (Table 5).

Table 4: Population differentiation regions from Table 1, which overlap with selected regions found previously, including A) F_{ST} regions from the 12k broiler dataset (Chapter 2, Tables 3 and 4), zero diversity regions in the B) 12K broiler (Chapter 3, Table 5), C) 42k broiler (Chapter 3, Table 2) and D) 600k broiler datasets (Chapter 3, Table 3) and E) the regions displaying significant asymptotic regression in the 600k broiler dataset (Chapter 3, Table 4).

Region No.	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNP count	QTLs	Broiler QTLs	Layer QTLs	Enrichment threshold	Genes	Overlaps
F_{ST}.2	2	27.14	27.50	0.35	70.61	70.71	0.10	10	1	0	1	9	1	B, C, E
F_{ST}.4	2	69.80	70.11	0.31	167.10	167.15	0.05	7	0	0	0	7	0	A
F_{ST}.7	3	51.28	51.63	0.36	125.84	127.26	1.41	11	1	0	0	10	12	A
F_{ST}.10	3	93.77	94.30	0.53	182.79	183.14	0.35	15	0	0	0	15	1	C
F_{ST}.11	4	56.90	57.23	0.33	123.95	124.66	0.70	12	1	1	0	8	0	A
F_{ST}.13	5	27.64	28.15	0.51	79.54	81.14	1.60	17	1	1	0	14	6	A
F_{ST}.15	5	46.56	46.88	0.33	105.54	105.70	0.16	11	2	1	0	8	0	C
F_{ST}.20	10	1.84	2.17	0.33	12.63	20.29	7.66	5	0	0	0	8	38	A
F_{ST}.22	10	4.40	4.87	0.47	31.72	32.44	0.72	16	0	0	0	14	6	A
F_{ST}.25	10	15.47	15.76	0.29	63.26	63.81	0.56	11	0	0	0	7	0	A
F_{ST}.26	11	1.82	2.23	0.41	5.36	6.10	0.74	14	1	0	0	11	22	A, B, E
F_{ST}.30	13	9.06	9.45	0.39	30.52	32.52	2.00	7	4	2	2	10	11	B, C
F_{ST}.32	20	6.00	6.47	0.47	25.05	25.59	0.54	15	1	0	0	14	4	B, C

Table 5: Zero diversity regions from Table 2, which overlap with selected regions found previously, including A) F_{ST} regions from the 12k broiler dataset (Chapter 2, Tables 3 and 4).

Region No.	Chr	Start (Mb)	End (Mb)	Size (Mb)	Start (cM)	End (cM)	Size (cM)	SNP count	QTLs	Broiler QTLs	Layer QTLs	Enrichment threshold	Genes	Overlaps
Div.1	3	49.73	50.05	0.33	123.31	123.65	0.34	9	0	0	0	8	3	A
Div.3	10	4.34	4.69	0.35	31.58	32.22	0.65	11	0	0	0	9	5	A

4.4.1 Discussion

Modern domestic chickens were split into two main roles approximately 70 years ago. Broilers, specialised for meat production, and layers, specialised for egg production. This split was accompanied by large phenotypic changes and the genomic signatures of selection are detectable by statistical techniques. In this study, 32 regions of high population differentiation were identified between a broiler and a layer dataset. Some of these regions may represent selection for the distinct roles of broilers and layers. Additionally, seven regions of zero diversity were found in the layer dataset. Three regions overlapped between the population differentiation and zero diversity regions. Fifteen regions were found which overlapped with selected regions found using high differentiation, low diversity and asymptotic regression in broiler lines.

4.4.2 Sliding Windows

The fixed size sliding window method and the method of determining regions means that the regions identified can vary in both region size and SNP count. Smaller regions with fewer SNPs will have a higher variance and are more likely to be at extreme values, and therefore are more likely to be a region of high population differentiation or zero diversity. We calculated how many equivalent regions of the same size and SNP count were zero diversity or above the population differentiation threshold. Several regions were the only high differentiation or zero diversity region of their size and SNP count. This includes regions F_{ST}.10 (Figure 1), F_{ST}.13 (Figure 4), F_{ST}.30 (Figure 3) and F_{ST}.32 from the population differentiation results and

regions Div.4 (Figure 5), Div.5 (Figure 5), and Div.7 from the zero diversity layer regions.

4.4.3 Composition of Regions

In the high population differentiation regions, 14 of the 32 regions contain QTL peak positions. Of these 14 regions, eight contained QTLs associated with broiler or layer traits. One region (F_{ST}.30, Figure 3) located on chromosome 13 contained both broiler and layer QTLs. Of the seven zero diversity regions found in the layer dataset, only two contain QTL peak positions, but neither included layer QTL. A number of highly differentiated regions contain SNPs fixed or nearly fixed for alternate alleles in broilers and layers, including regions F_{ST}.10, F_{ST}.12, F_{ST}.13, F_{ST}.21, F_{ST}.22, F_{ST}.23, F_{ST}.30, F_{ST}.31, Div.2, Div.3 and Div.5 (Figures 1-6). This suggests that selection of both broilers and layers is driving differentiation in these regions.

4.4.4 Candidate selected regions

Several candidate regions for selection were found in this study, many of which overlapped with regions from the broiler population differentiation, zero diversity or regression regions. Region F_{ST}.2 located on chromosome 2 overlapped with regions in the 12k and 42k (broiler) zero diversity regions and 600k (broiler) regression regions. One layer QTL related to egg weight was found in this region, as well as one gene (*ER81*), which has not previously been associated with broiler or layer traits (Table S12, Region 2). 12 genes were found within 1Mb of this region, but none were functional candidates for broiler traits. Region F_{ST}.26 on chromosome 11 overlapped with 600k regression regions, 12k diversity and 12k F_{ST} broiler regions.

Twenty-two genes were found in this region, as well as one QTL related to fear response (Schütz et al., 2004). 35 genes were found within 1Mb of the region but none are associated with broiler traits.

Region F_{ST}.30 (Figure 3) located on chromosome 13 contained four QTL peak positions (Table S12, Region 30). Two QTL were related to broiler traits, and include body weight and drumstick weight QTL, while the other two QTL were related to layer traits, such as age at first egg. Eleven genes were found in this region and 64 genes were found within 1Mb of the region. This region overlaps with 12k and 42k (broiler) zero diversity regions. The high differentiation here suggests that the alternative allele is selected in the layers. Of the 7 SNPs in this region, 3 are fixed for the alternate allele in broilers and layers and 2 more are fixed in broilers and nearly fixed (allele frequency is 0.986) for the alternate allele in layers. This suggests that selection is occurring in both broilers and layers in this region. Of 299 equivalent regions of its size and SNP count, this was the only highly differentiated region.

A number of identified regions contain candidate genes relating to reproduction traits. Region F_{ST}.1 located on chromosome 1 contains the *SMO* gene (Table S12, Region 1). This receptor is associated with hedgehog proteins, including *sonic hedgehog* which in encoded by *SHH*. These proteins are known to have important roles in embryogenesis. *Sonic hedgehog* plays key roles in limb digit growth (Riddle et al., 1993) and the organisation of the brain (Herzog et al., 2003) and spinal cord (Litingtung and Chiang, 2000). Region F_{ST}.20 (Figure 2) located on chromosome 10 overlaps with the *ULK3* gene. This gene is also known to interact with *sonic hedgehog*. Another gene overlapping with region F_{ST}.20 is *ISLR2* and is associated with neural development (Table S12, Region 20). Finally, region F_{ST}.23/Div.5

(Figure 5) overlaps with the *FGF7* gene, which has been shown to be important in lung development in rats (Shiratori et al., 1996) and the late stages of organ development in mice (Nguyen et al., 1996). One of the SNPs in this region is located within this gene, and is fixed for alternative alleles in the broilers and layers. Therefore, this gene may be selected in layers.

Several regions overlapped with signatures of selection identified in previous studies. Region F_{ST}.32 located on chromosome 20 overlapped with a region of low diversity in layer chickens (Qanbari et al., 2012). This region contains four genes and one QTL relating to lactate level (Table S11 and S12, Region 32). Five regions (F_{ST}.14, F_{ST}.16, F_{ST}.18, F_{ST}.30 (Figure 3) and Div.5 (Figure 5)) overlapped with low diversity regions identified in several commercial and non-commercial lines (Rubin et al., 2010).

4.4.5 Time of Selection

A total of seven regions were found displaying high differentiation between the layer and broiler datasets and between the 12k broiler datasets (Chapter 2, Tables 3 and 4). The fact that these regions are differentiated both between broiler lines and broiler and layers suggests that these regions may represent selection that occurred during breed separation rather than selection for broiler or layer specialisation. Two of these regions (F_{ST}.11 and F_{ST}.13) contain QTL relating to broiler traits. The remaining regions may indicate selection which occurred later during line specialisation.

A number of the same chicken breeds were used in the development of both the broiler and layer lines. Brown egg layer chickens were based on a number of dual purpose breeds selected for both meat and egg qualities, including White Plymouth

Rock and Rhode Island Red (Muir et al., 2008). Broilers are produced using specialised crosses, with the male lines selected principally for growth and the dam lines selected for both growth and reproductive traits. The male lines were based on British Cornish Indian Game breeds while the dams are based on dual purpose breeds, many of which were used in layer development, including White Plymouth Rock and Barred Plymouth Rock (Muir et al., 2008).This may explain why a number of regions were differentiated between broiler lines and between broiler and layer lines.

4.4.6 Future Work

The population differentiation analysis uses the outlier method to identify putative regions under selection. While this method has been used in several previous studies (Akey et al., 2002, Akey et al., 2010, Wilkinson et al., 2013), it does not include a confidence interval or p-value. Chapter 2 included a circular chromosome bound permutation method to generate a null distribution for each F_{ST} window (See section 2.2.6 for more details). However, this method requires at least three different populations to be compared in a pairwise F_{ST} calculation and was therefore not appropriate for use here. Additionally, in previous chapters, close family members (sibs, half sibs and parents) were removed by the quality control procedures. However, pedigree information was not available for the layer dataset, so it could not be included. In future studies, this problem could be addressed by using programs such as GCTA (Yang et al., 2011) to estimate and remove close relatives. It would also be useful to investigate non-synonymous changes within the candidate genes described above. This could be achieved using whole genome sequence data in these lines, as the technology is become more affordable over time. Finally, focusing on

regions of zero diversity in the layer dataset will prevent us from detecting sweeps where the allele is not fixed.

4.4.7 Conclusions

A total of 32 regions were identified in a layer and broiler line that had extreme values of F_{ST}. An additional seven regions were identified that displayed zero diversity in the layer dataset. Several QTLs relating to layer and broiler traits were found in these regions. Several high differentiation regions overlapped with 42k broiler zero diversity regions or 42k layer zero diversity regions, and were found to be fixed for alternative SNP alleles in the broilers and layers. Seven high differentiation regions overlapped with high differentiation regions found between broiler datasets (Chapter 2), suggesting these regions represent selection that occurred during breed separation. This may have occurred as some of the same breeds, such as White Plymouth Rock, were used as the basis of some layer lines and broiler dam lines. The remaining regions may represent more recent selection which occurred during line specialisation.

CHAPTER FIVE

General Discussion

5.1 Thesis aims and objective overview

The aim of this thesis was to investigate the effects of artificial selection on the genome of domesticated chickens. Genomic regions displaying signatures of selection are likely to be associated with commercial traits and are therefore of interest to breeders. Selection was investigated using three main statistical tests, which examine population differentiation and local reductions in diversity. These tests uncovered a number of putative selected regions, some of which include candidate genes associated with broiler or layer development.

5.2 Conclusions and implications of findings

5.2.1 Selected regions

A number of selected regions were found in multiple datasets (Figure 1) using different statistical techniques (Figure 2). A region on chromosome 2 located at approximately 27Mb was found as region of low diversity in the 12k and 42k broiler datasets, a region of significant asymptotic regression in the 600k broiler dataset and finally as a region of high differentiation between 42k broiler and layer datasets. This region contained an egg weight QTL association related to layer traits and one gene. This region may be selected in both broilers and layers, as selection signatures were identified in both groups.

Figure 1: Venn diagram describing the number of distinct regions found in each chapter and the number of these regions which overlap between the chapters. Regions found multiple times in the same chapter within different datasets are counted once.

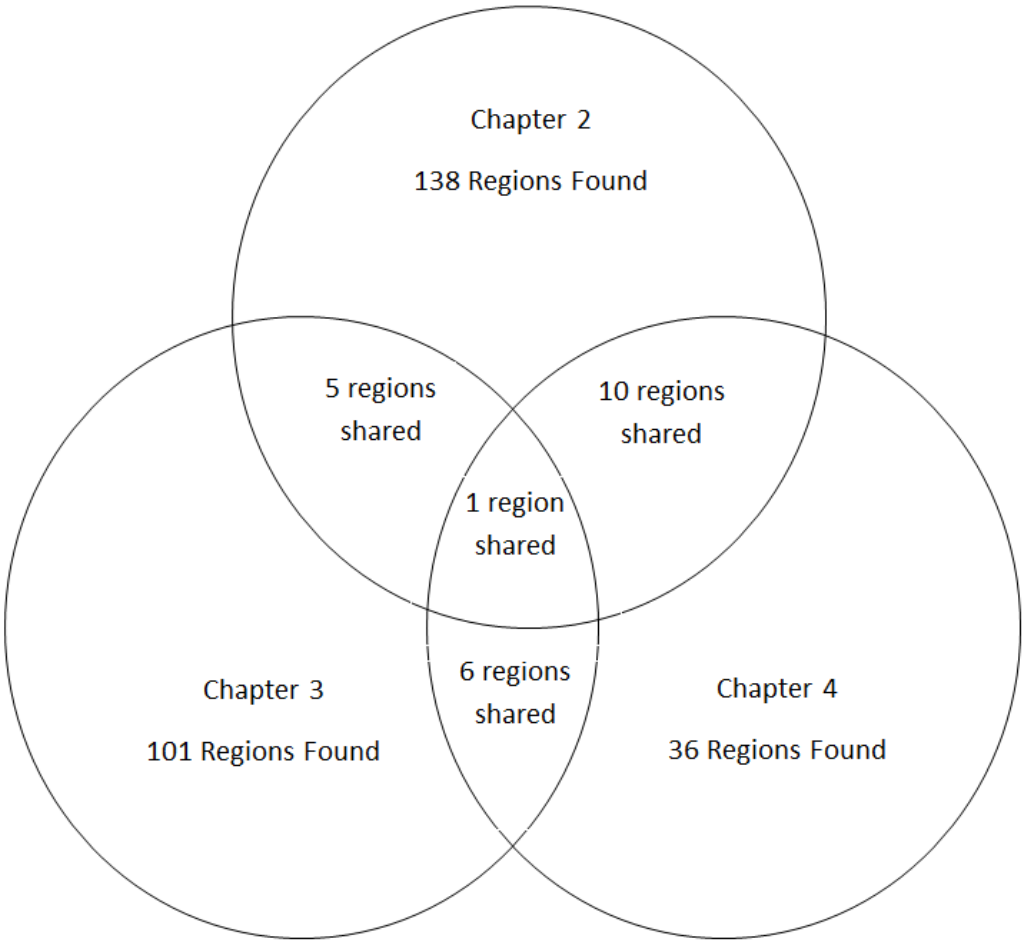
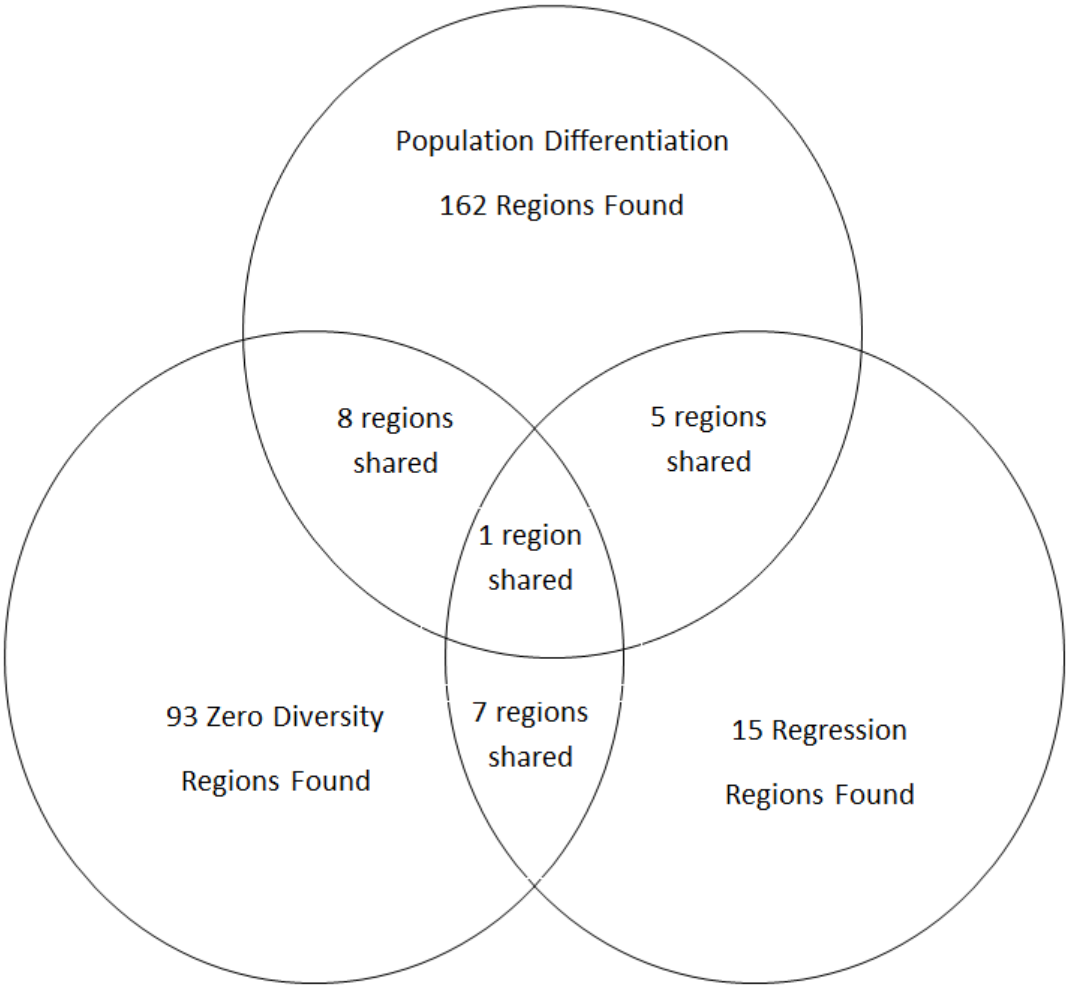


Figure 2: Venn diagram describing the number of distinct regions found using each method and the number of these regions found in more than one method. Regions found multiple times using the same method within different datasets are counted once.



Other candidate regions in the layer dataset include a region on chromosome 1 located at approximately 0.5Mb which contains a *smoothened, frizzled class receptor gene* (*SMO*) (Figure 3). Another high differentiation region on chromosome 10 located at approximately 2Mb contains *unc-51 like kinase 3* (*ULK3*) (Figure 4). This region is also highly differentiated between broiler lines which could represent selection which occurred during the creation of the lines. Both *SMO* and *ULK3* are known to interact with hedgehog proteins, which are involved in embryogenesis (see section 5.2.2 for more details).

Figure 3: Genes found in region F_{ST}.1 of high differentiation between broilers and layers located on chromosome 1 between 0.472Mb and 0.803Mb, highlighted in the red box. This includes the candidate gene *SMO*, which is known to interact with hedgehog proteins. Gene location data is taken from Ensembl (http://www.ensembl.org/).

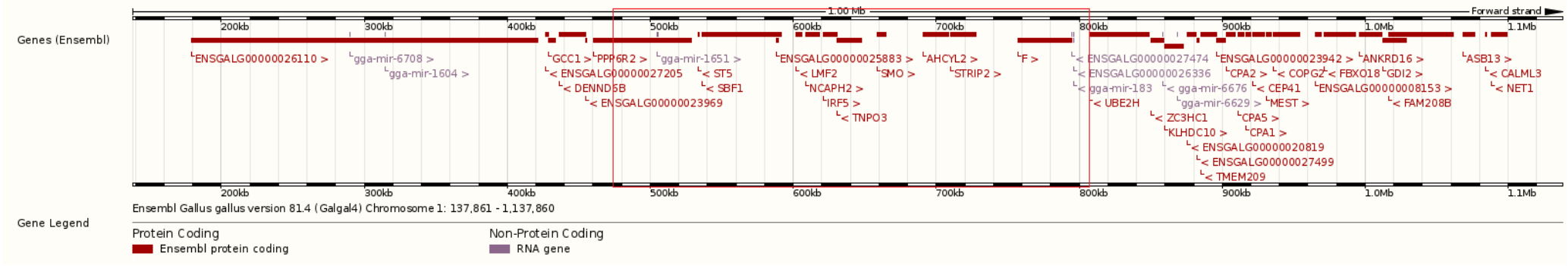
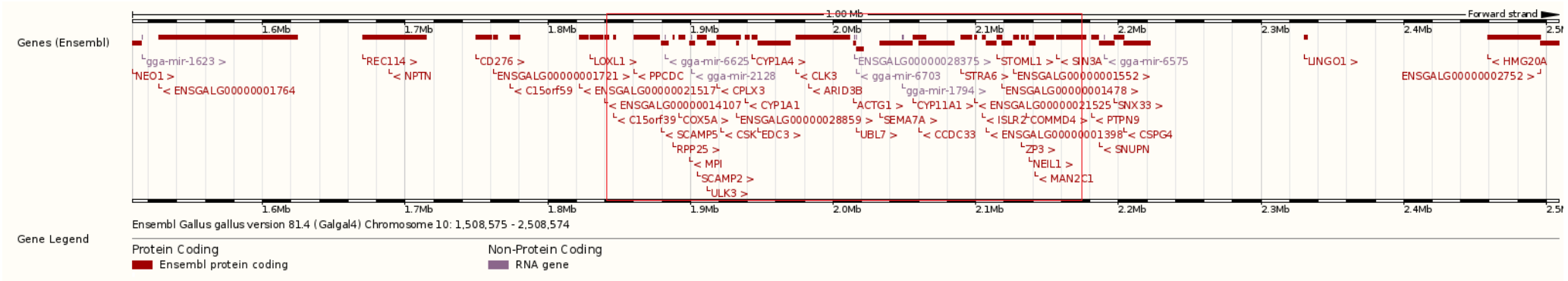
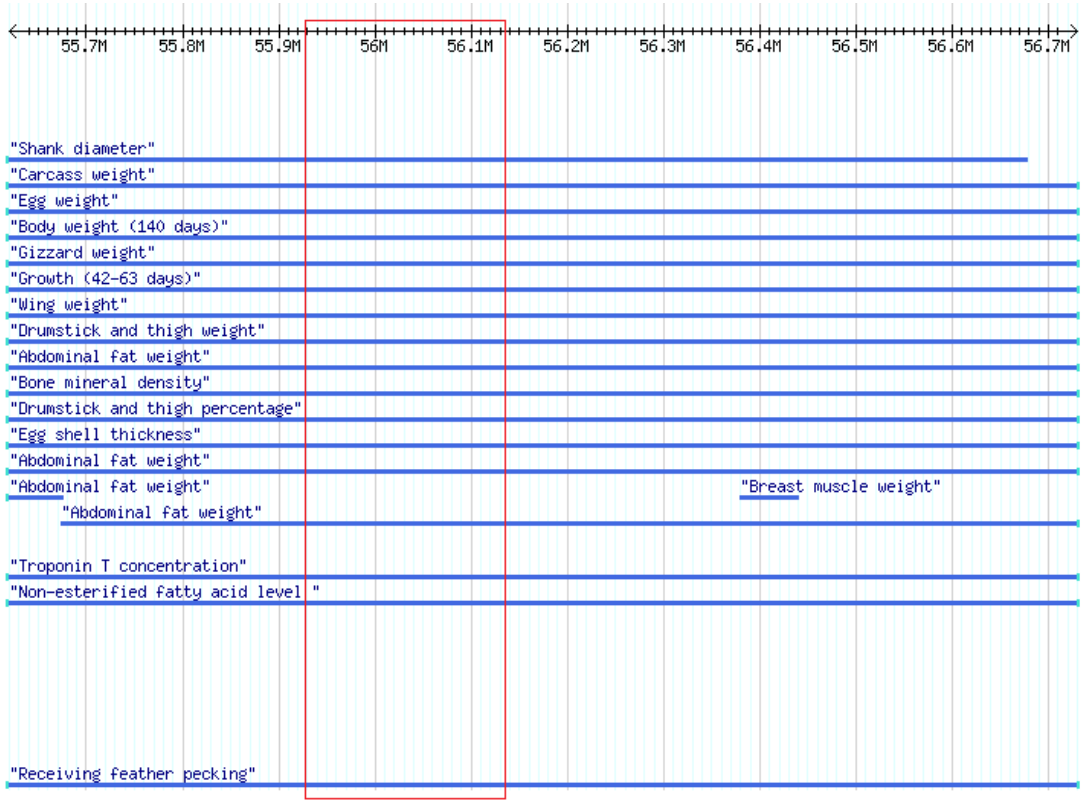


Figure 4: Genes found in region F_{ST}.20 of high differentiation between broilers and layers located on chromosome 10 between 1.843Mb and 2.173Mb, highlighted in the red box. This includes the candidate gene *ULK3*, which is known to interact with hedgehog proteins. Gene location data is taken from Ensembl (http://www.ensembl.org/).



A region was found on chromosome 5 at approximately 55Mb, which was found as a region of low diversity in the 42k broiler dataset, a region of asymptotic regression in the 600k broiler dataset and a highly differentiated region between broiler lines 1, 2, 3, 4 and 6 (Figure 5). This region was not highly differentiated between broiler and layer lines (chapter 4), which may suggest this selection occurred after breed separation within individual broiler lines. This region contains a body weight QTL which is related to broiler traits. Five genes were found in this region, but were not obvious candidates for broiler traits.

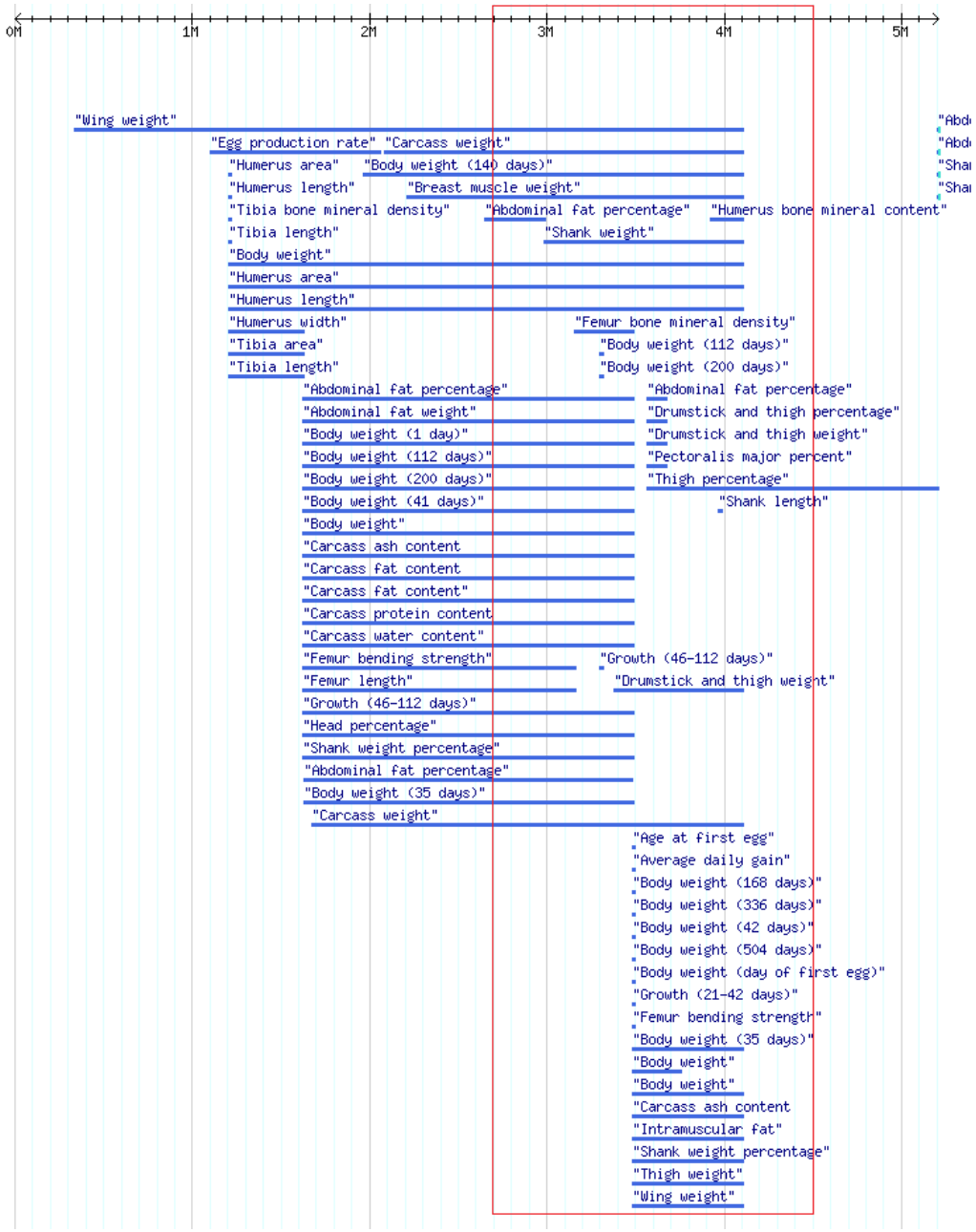
Figure 5: QTL spans from the chicken QTL database (Hu et al., 2013) found on chromosome 5 between 55.928Mb and 56.136Mb. Region TReg.9 displayed significant asymptotic regression in the 600k broiler dataset is displayed as a red box (found in chapter 3). The Body Weight (140 days) QTLs peak position is found within this region. The other QTLs displayed have spans which overlap this region but not a peak position.



A region on chromosome 27 at approximately 3Mb was found as a region of high differentiation between broiler lines (Figure 6). This is an interesting region as it was

found in seven broiler lines, including 1, 4, 5, 6, 7 and 8, but not in the broiler vs layer high differentiation regions (Chapter 4). This may suggest this selection occurred in different broiler lines after the broilers and layers were separated. A number of interesting candidate genes were found in this region, related to insulin like growth factor proteins, discussed below (see section 5.2.2 for more details). Thirty nine QTL relating to broiler traits were found in this region, including body weight, growth rate and abdominal fat weight. However, this region was found in the low density 12k broiler dataset. The resolution available in this dataset is low. Therefore the size of the region is large (approximately 1.78Mb), which reduces the chance of confidently identifying the selected gene.

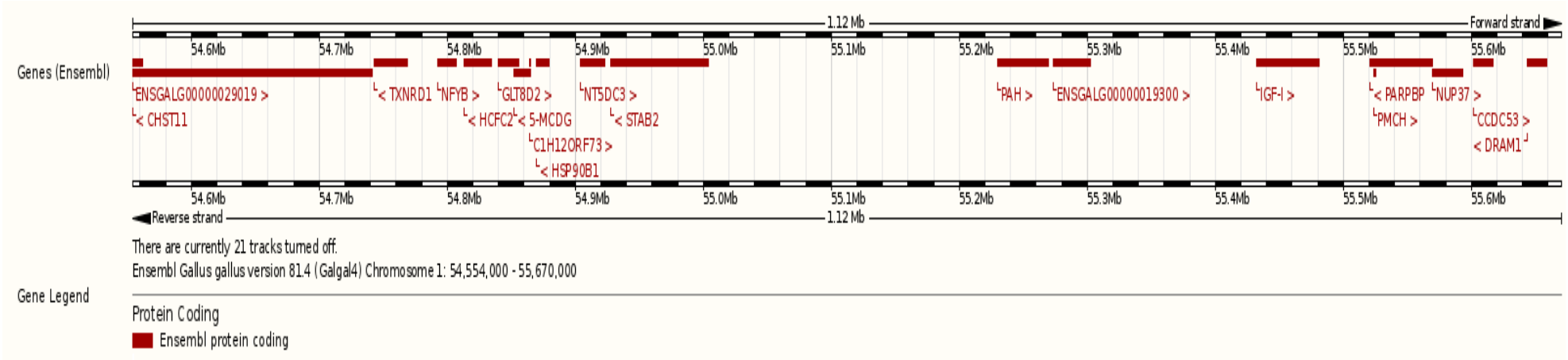
Figure 6: Production QTL spans from the chicken QTL database (Hu et al., 2013) found on chromosome 27 between 2.72Mb and 4.50Mb. The highly differentiated region 50 (found in multiple lines) is displayed as a red box (chapter 2). This region may include more QTLs than mentioned in the text as the whole QTL span is displayed rather than a peak position. This region is very QTL rich so only production QTLs are included in this figure.



Some regions found in this thesis have been identified in previous studies. A region of significant asymptotic regression on chromosome 1 (Figure 7) located at

approximately 55Mb (See chapter 3 table 4 for more details) was also identified using a pooled heterozygosity method in layer lines (Qanbari et al., 2012). Additionally, it was identified as a zero diversity region in the 12k and 42k datasets but not in the layer dataset in chapter 4. Therefore, broilers are likely to be driving the selection in the region. This region also contains insulin-like growth factor 1 *(IGF1),* a gene often associated with broiler traits (see section 5.2.2. for more details).

Figure 7: Genes found in a region of significant asymptotic regression located on chromosome 1 between 54.554Mb and 55.670Mb. The region TReg.2 runs for the complete length of the figure and includes the candidate gene *IGF1*. Gene location data is taken from Ensembl (http://www.ensembl.org/).



A region on chromosome 11 located between 1 and 4Mb was identified displaying significant asymptotic regression in the 600k dataset, high differentiation between 42k broiler and layer datasets and zero diversity in the 12k and 42k broiler datasets. Therefore, this is the only region to be found in all 3 chapters. This region was also identified in lines divergently selected for abdominal fat content (Zhang et al., 2012a) as well as in low diversity regions in two sets of pooled commercial and non-commercial chicken lines (Rubin et al., 2010, Elferink et al., 2012). This region contains 70 genes and two QTL relating to abdominal fat content, which may suggest that broiler selection is present in this region.

5.2.2 Candidate genes

A number of candidate genes were found within selected regions. This includes several genes relating to insulin like growth factors (*IGF*). Two were found in region 50 located on chromosome 27 in the high differentiation regions between broiler lines (see section 2.4.3 for more details), including genes which encode insulin-like growth factor 2 mRNA binding protein 1 *(IGF2BP1)* and insulin-like growth factor binding protein 4 (*IGFBP4*), which bind to and alters IGFs interaction with cell receptors. *Growth hormone 1 (GH1*) was also found within 1Mb of this region. Additionally, the *insulin-like growth factor 1* (*IGF1*) gene was identified in a region of significant asymptotic regression in the 600k broiler dataset and in zero diversity regions in the 12k and 42k broiler datasets (see section 3.4.4 for more details). There is no evidence of selection for *IGF1* in the layer line (Chapter 4).

Insulin-like growth factor 1 (IGF-1) is a protein coded by the *IGF1* gene. IGF-1 is similar in structure to insulin and has been shown to play a role in growth. In humans, it is produced mainly in the liver and production is stimulated by growth hormone (GH). IGF-1 binds to Insulin-like growth factor 1 receptor (IGF1R). The offspring of mice with a targeted mutation in *IGF1* are infertile dwarfs, approximately 30% of normal size (Baker et al., 1996). Another study found that mice with a mutation in *IGF1* display a growth deficiency, with some dwarfs dying soon after birth (Liu et al., 1993). Additionally, mice with mutations in *IGF1R* suffer from a larger growth deficiency and die soon after birth.

IGF1 has been previously implicated by other selection studies in a number of species. A study in dogs found that a specific haplotype in *IGF1* was common in all small dog breeds (Sutter et al., 2007). *IGF1* is associated with tissue maintenance and milk production in dairy cattle (Flori et al., 2009) and is associated with feed intake, growth and carcass traits in beef cattle (Johnston et al., 2001, Wood et al., 2004). In chickens, *IGF1* has been implicated by a number of selection studies. This includes low diversity regions found in pooled data from a several commercial and non-commercial populations (Rubin et al., 2010), in pooled data in laying chickens (Qanbari et al., 2012) and in two chicken populations divergently selected for abdominal fat content (Zhang et al., 2012a). Therefore, genes relating to *IGFs* are likely to be involved in broiler selection.

In the layer dataset, two genes were found which relate to hedgehog proteins. The first was a *smoothened, frizzled class receptor* gene (*SMO*) located on chromosome 1 and the second was the *unc-51 like kinase 3* (*ULK3*) gene located on chromosome

10. Hedgehog proteins, especially sonic hedgehog, are associated with embryogenesis, and could be selected in layers.

The *hedgehog* genes (*Hh*) were first identified in *drosophila melanogaster* in studies to detect genes which affect the segmentation pattern in *drosophila* embryos (Nusslein-Volhard and Wieschaus, 1980). The sonic hedgehog protein encoded by the *sonic hedgehog* gene (*SHH*) is involved in central nervous system development (Litingtung and Chiang, 2000), tooth development (Thesleff, 2003) and limb digit growth in vertebrates (Riddle et al., 1993). Limb digit growth is particularly well studied in chickens as eggs allow direct access to the embryo during development (Yang et al., 1997). Other hedgehog genes include *Indian hedgehog* (*IHH*) which is involved in cartilage differentiation (Vortkamp et al., 1996) and *Desert hedgehog* (*DHH*) is involved in cell signalling, spermatogenesis (Clark et al., 2000) and testes development in mice (Bitgood et al., 1996). As hedgehog genes are heavily involved in the development of the embryo, it is possible they are under selection in layer chicken lines.5.2.3 Findings in relation to broiler and layer development

A total of seven regions were found to be highly differentiated between broiler lines as well as between the broiler and layer lines. These include regions F_{ST}.4, F_{ST}.7, F_{ST}.11, F_{ST}.13, F_{ST}.20, F_{ST}.22 and F_{ST}.25 located on chromosomes 2, 3, 4, 5 and 10 (See chapter 4, Table 4 for more details). These regions may be the product of selection which occurred during line separation, prior to specialization of broilers and layers. During the development of broiler and layer lines, a number of the same breeds were used as the basis of both brown egg layers and broiler dam lines, including the White Plymouth Rock breed (Muir et al., 2008). The use of certain breeds in both broiler and layer development could result in regions which are

differentiated between both broiler lines and broiler and layer lines. The remaining 25 regions are more likely to represent selection which occurred later, after the lines were separated.

5.2.3 Candidate genes found in previous studies

Candidate genes under selection have been found by a number of previous selection signature studies using a number of commercial and non-commercial chicken populations (See section 1.4.2 for more details). Some of these genes were also found in this study, including *IGF1*, *IGF2BP1*, *GH* and *GPD2*. However, several candidate genes, including *TSHR*, *BCDO2* and *TRPC4* were not found in any of the broiler or layer datasets used in this study. In order to investigate reasons why these genes are not present in the results, the diversity values in the location of these genes was examined in the 600k broiler dataset, as this dataset offers the greatest resolution.

The *TSHR* gene located on chromosome 5 was found in pooled populations of broiler, layers and red junglefowl using a z-transformed heterozygosity (ZH_P) method to detect local reduction in diversity (Rubin et al, 2010). This gene is associated with metabolic regulation and reproduction and may be related to the change to non-seasonal reproduction found in domesticated commercial animals compared to seasonal reproduction in wild animals (Hanon et al, 2008). In the 600k broiler dataset, two SNPs were found within this gene with a diversity of zero and three more were found with a diversity score of approximately 0.0075. When these SNPs were averaged into sliding windows, only one window within this gene was found to have a diversity of zero while another had a very low diversity score of

0.0037. In this case, the criterion to create a region may be too stringent, as two unique windows of zero diversity were required to form a zero diversity region. Therefore, this gene was not detected in this study. Alternatively, relaxing the criteria from zero diversity to a low diversity, (for example, where diversity is less than 0.004) would also allow this gene to be detected.

The second gene investigated was *BCDO2*, located on chromosome 24. This gene is related to yellow skin colour (Eriksson et al, 2008) and several studies have used it as a test to prove that the method used can find established selective sweeps (Rubin et al, 2010, Elferink et al, 2012). In the 600k broiler results, two SNPs are found with a diversity of zero and four other SNPs are present with diversity less than 0.008. Two sliding windows which overlap with this gene have low diversity scores of 0.028 and 0.040 and contain 10 SNPs each. This region was not detected as a zero diversity window as there were many SNPs present within the sliding windows, some of which were segregating. Relaxing the zero diversity criteria may allow this gene to be detected in future studies.

TRPC4 was detected as a candidate gene on chromosome 1 in two lines divergently selected for abdominal fat content (Zhang et al, 2012a). The region containing this gene had strong EHH signals in the fat line, as well as large allele frequency differences between the two lines suggesting that this gene may be related to increased abdominal fat. Eight SNPs within this gene had a diversity score of zero. However, there are a large number of SNPs located within this gene and many of them are segregating. Therefore, the diversity scores for the overlapping sliding windows are relatively high, ranging from 0.09 to 0.32. It is unlikely that this region

would be detected even if the zero diversity criterion was relaxed to a low diversity criterion.

HNF4G was detected in regions of low diversity in 67 chicken breeds, also using the ZH_P method (Elferink et al, 2012). This gene is located on chromosome 2 and is associated with larger body weights in mice (Gerdin et al, 2006). This gene also contains some zero diversity SNPs in the 600k broiler dataset but many of the surrounding SNPs diversity are at approximately 0.2. This creates sliding window diversity averages of approximately 0.1. This gene is also unlikely to be detected in a low diversity study using this dataset. It is possible that this gene was previously detected due to the large number of pooled lines used in the previous study (Elferink et al, 2012).

Finally, the *CAPN1* gene was detected as a candidate gene in a region of high population differentiation between brown and white layer chickens (Gholami et al, 2014). This gene has previously been associated with meat quality (Zhang et al, 2008) and is located on chromosome 3. The SNP diversity in the 600k broiler dataset present in this gene is relatively high with the majority above 0.3. As this gene was found within layers, 42k brown layer dataset was also investigated in the region of this gene. The diversity in this area was found to be approximately 0.3. It is possible that this gene was not detected due to the different datasets used in this study.

This investigation suggests that in some cases, the criteria for creating a region may be too stringent. Currently, two sliding windows of zero diversity are required to be detected as a zero diversity region. This was to prevent the majority of regions consisting of sliding windows with very few SNPs which will show greater variance,

and possibly more extreme values. However, this rule also prevented the detection of an established sweep at the *TSHR* gene (Rubin et al, 2010). Additionally, while only using zero diversity regions may reduce the number of regions found to the most extreme examples, it also means that some established sweeps may be missed. This includes the region at the *BCDO2* gene which may have been located if this criteria was relaxed. However, if the criterion is relaxed, then many more regions will be detected, and therefore more regions which have a low diversity value due to demographic events rather than selection. The effect of relaxing this criterion is investigated in section 5.3.3.

5.2.4 Methodological issues

A circular chromosome bound permutation method was developed to identify regions of high differentiation between nine broiler lines, based on Cabrera et al. (2012). This generated a null distribution for each window and allowed statistically extreme values of F_{ST} to be identified. This method prevented bias towards windows with fewer SNPs and therefore more extreme values of F_{ST}, which is present in the outlier method (Akey et al., 2002). However, one limitation of this method is that is requires pairwise population differentiation to be calculated for at least three different populations. Therefore, it could not be used in chapter 4 to identify regions of high differentiation between the broiler and layer lines. In this case, the outlier method was used to identify high differentiation regions in the upper tails of the empirical distribution of F_{ST} values. These high differentiation regions are more likely to be biased towards windows with fewer SNPs and consequently may contain more windows which are highly differentiated by chance.

Regions of low diversity can be caused by selection. The selected allele spreads throughout the population and becomes fixed, and the hitchhiking effect also reduces variation at neutral sites linked to the selected site. This causes an area of low variation around the selected gene. However, certain demographic events, such as population bottlenecks or founder events, can also create reduce diversity in the genome. It can be difficult to differentiate between these different events. The SNP panels used to generate data analysed in this thesis were designed to include markers segregating in commercial breeds and therefore zero diversity regions should be rare. In order to understand how rare a zero diversity region of a specific size and SNP count is, we developed a method where all possible regions of its size and SNP count were investigated. The proportion of these with a diversity of zero was calculated. Regions with a small proportion of zero diversity regions are less likely to have occurred by chance. However, this method cannot confirm whether this was caused by selection or demographic events.

5.3 Future work

As discussed in the introduction, there are many tests available for detecting selection. These tests fall roughly into four distinct categories; differentiation between populations, local reductions in diversity, changes to the frequency spectrum and linkage disequilibrium based analysis. In this thesis, three tests from two of the categories were investigated and changes to the frequency spectrum and linkage disequilibrium were not investigated. Some of these tests could be run on the datasets in future. In particular the high density 600k broiler dataset, as it offers the greatest resolution.

Some tests could not be used due to the format of the SNP data. The SNP files provided do not include specific nucleotides. Instead, a numerical code is used. A "top allele" was defined during genotyping. If the SNP is homozygous for the top allele, this is represented by a 2. If the SNP is homozygous for the alternate allele, it is represented by a 0. Finally, if a SNP is heterozygous for both alleles, this was represented by a 1. We did not know the identity of either allele in any SNP. Additionally, there were no SNP IDs present in any dataset except from the 600k line 3 broiler line. This was not a limitation for the diversity and differentiation tests used as is it still possible to identify regions of low diversity and high population differentiation from these codes. However, if more information was present, it would allow additional analysis to be performed. For example, the iHS test requires the identification of the ancestral allele. This type of information can be found in dbSNP (Sherry et al., 2001). However, the numerical SNP codes made it impossible to differentiate between the ancestral and derived alleles at each SNP. It is also possible that in datasets that were genotyped several years apart, such as the 12k and 42k datasets, the top allele may have switched in some SNPs. This means that where previously a homozygous allele would have been represented by a 2 in these SNPs, it would now be represented by a 0. This causes problems when investigating allele frequency changes over time between the two datasets.

Other information is missing which may have aided with the identification of signatures of selection. No phenotype information for the lines was available to us. This made it impossible to focus on specific traits that the individual lines were selected for. Finally, of the four SNP datasets used in this thesis, three are of relatively low density. Statistical tests can detect selection within low density

datasets, but it is more difficult to narrow down the region to confidently identify candidate genes. We determined the sliding window size for each dataset based on an average of approximately 10 SNPs per window so lower density datasets had much larger window sizes. Therefore, the available resolution in these datasets is lower and the region sizes are larger which makes it more difficult to confidently identify the selected gene within a region. This effect can be seen in chapter 3, where the average zero diversity region size in the 12k line 3 broiler dataset is 1.25Mb (Chapter 3, Table 1), the average zero diversity region size in the 42k broiler dataset is 0.52Mb (Chapter 3, Table 2) and the average zero diversity region size in the 600k broiler dataset is 0.03Mb (Chapter 3, Table 3). However, if smaller window sizes were used, most windows in the low density datasets would contain very few SNPs.

5.3.2 Identifying and analysing additional candidate genes

In this thesis, a number of candidate genes found in highly differentiated and zero diversity regions were identified in several broiler datasets and one brown egg layer dataset. Some of these regions, including *IGF1* have been previously discovered in other studies (Rubin et al, 2010, Elferink et al, 2012, Zhang et al, 2012a, Fan et al, 2013) and are therefore likely to be under selection. However, a number of regions have no obvious candidate genes present which are related to broiler or layer traits or additional evidence of selection from the literature. These regions will need additional research to determine whether they are under selection, and to identify the gene responsible.

One method would be to utilise whole genome sequencing. This technology is constantly maturing, becoming more accurate and inexpensive (Ng and Kirkness, 2010) and would produce a whole genome sequence for each genotyped bird which

would improve the results of this study in a number of ways. Firstly, the availability of the complete genome sequence would allow more statistical methods to be used and therefore more regions to be identified. Linkage disequilibrium methods such as EHH (Sabeti et al., 2002, Voight et al., 2006) and frequency spectrum based methods (Fay and Wu, 2000) could be used as the identity of alleles would be known instead of the recoded SNPs present in the SNP panels. Additionally, the linkage disequilibrium present in the population could be analysed across the genome. Secondly, whole genome sequences would allow the current identified regions to be further investigated with higher accuracy. For example, it would be much easier to locate exactly where a region of low diversity begins and end, rather than estimating the location using sliding windows. This would allow candidate genes to be more easily identified. Finally, whole genome sequencing would allow the candidate genes to be investigated in much greater depth. Gene sequences could be downloaded from an online source (such as Ensembl or NCBI) and then compared to the copy of the gene found in the chicken population for any non-synonymous changes. These non-synonymous changes may indicate causal candidate mutations in these genes which are responsible for the selective sweeps.

Another option would be to utilise alternative sequencing methods, such as RNA-seq, which can be used to measure RNA quantity from a genome at a particular moment in time (Nagalakshmi et al, 2008). This could be used to investigate differences in gene expression in the muscles of different populations of chickens. For example, this could be used in the nine lines of broiler chickens used in chapter 2 or between the line 3 600k broiler dataset used in chapter 3 and an ancestral population of red junglefowl. This technique would allow an understanding of the

differences in gene expression in these regions and therefore important differences between the lines at a particular time.

5.3.3 Effects of using zero diversity regions

In chapters 3 and 4, a number of zero diversity regions were identified in broilers and layers. However a diversity of zero is a very stringent criterion. This approach will not detect sweeps which have not reached fixation and therefore may not be the best method of detecting recent selective sweeps. For example, some sweeps may not have had sufficient time to reach fixation. Additionally, recombination or synonymous mutation can increase diversity in a region which would prevent the detection of a region. In order to capture these sweeps, the zero diversity criterion would need to be relaxed.

To investigate the number of regions that could be found in the 600k dataset if the zero diversity criteria was relaxed, regions were re-created using sliding windows with diversity scores less than or equal to 0.001, 0.005, 0.01 and 0.05. When using the 0.001 diversity limit, 15 regions of low diversity were found. When the 0.005 diversity limit was used, 449 putative selected regions were found. With the 0.01 diversity limit, 625 putative selected regions were found. Finally, when using the 0.05 diversity limit, 1340 regions were found.

Seven regions were found in the 600k broiler dataset when only using windows of zero diversity, while 15 were found using the diversity limit of 0.001. This relatively small increase in region number suggests that in future studies, using a low diversity limit in high density datasets may allow the detection of additional selected regions while introducing relatively few false positives into the results. If the diversity limit

is increased past 0.001, then a large number of regions are detected. At 0.005 diversity, a relatively small increase, the number of regions detected increases from 7 to 449 and increases to 1340 when 0.05 is used. While a number of these new regions will represent selection on the genome, a large number are likely to be false positives. This may include regions which have become fixed by chance due to drift, inbreeding or demographic events. Therefore in future work, it will be important not to relax the diversity criterion excessively to strike a balance between detecting selected regions and introducing a large amount of non-selected regions into the results. The exact criterion may be dependent on the SNP density in the dataset used. An alternative method would be to use this diversity criterion alongside another method, such as the regression method (Wiener and Pong-Wong, 2011) to idenitify signatures of selection (see section 3.2.8 for more details).

5.4 Conclusion

In conclusion, this thesis identifies a number of signatures of selection in both broiler and layer chickens. A circular chromosome-bound permutation method was developed to identify regions of high differentiation between broiler datasets. Regions of low diversity were investigated in 3 broiler datasets and regions of significant asymptotic regression were found in the high density, 600k broiler dataset. A number of candidate genes were identified, including *IGF1*, which are related to broiler traits. Finally, regions of high population differentiation were identified between a broiler and layer line, as well as low diversity regions in the layer lines. Several genes related to hedgehog protein were located within these regions, which are known to be involved in embryogenesis and could be selected in layers. Seven regions were found to be differentiated between broilers and layers as

well as between broiler lines, which could indicate selection which occurred during the separation of broiler and layer lines. With additional information for each of the datasets, more statistical tests could be utilised, including tests based on linkage disequilibrium and the frequency spectrum. This could confirm current selected regions, as well as identify new regions undetectable in the current methods.

Bibliography

Ai, H., Huang, L. and Ren, J. (2013) Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers, *PLoS ONE*, **8**(2), e56001.

Akey, J. M., Ruhe, A. L., Akey, D. T., Wong, A. K., Connelly, C. F., Madeoy, J., Nicholas, T. J. and Neff, M. W. (2010) Tracking footprints of artificial selection in the dog genome, *Proceedings of the National Academy of Sciences*, **107**(3), 1160-1165.

Akey, J. M., Zhang, G., Zhang, K., Jin, L. and Shriver, M. D. (2002) Interrogating a High-Density SNP Map for Signatures of Natural Selection, *Genome Research*, **12**(12), 1805-1814.

Allison, A. C. (1954) Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection, *BMJ*, **1**(4857), 290-294.

Ambo, M., Moura, A. S. A. M. T., Ledur, M. C., Pinto, L. F. B., Baron, E. E., Ruy, D. C., Nones, K., Campos, R. L. R., Boschiero, C., Burt, D. W. and Coutinho, L. L. (2009) Quantitative trait loci for performance traits in a broiler × layer cross, *Animal Genetics*, **40**(2), 200-208.

Amills, M., Jimenez, N., Villalba, D., Tor, M., Molina, E., Cubilo, D., Marcos, C., Francesch, A., Sanchez, A. and Estany, J. (2003) Identification of three single nucleotide polymorphisms in the chicken insulin-like growth factor 1 and 2 genes and their associations with growth and feeding traits, *Poultry Science*, **82**(10), 1485-1493.

Andreescu, C., Avendano, S., Brown, S. R., Hassen, A., Lamont, S. J. and Dekkers, J. C. M. (2007) Linkage Disequilibrium in Related Breeding Lines of Chickens, *Genetics*, **177**(4), 2161-2169.

Ankra-Badu, G., Shriner, D., Le Bihan-Duval, E., Mignon-Grasteau, S., Pitel, F., Beaumont, C., Duclos, M., Simon, J., Porter, T., Vignal, A., Cogburn, L., Allison, D., Yi, N. and Aggrey, S. (2010) Mapping main, epistatic and sex-specific QTL for body composition in a chicken population divergently selected for low or high growth rate, *BMC Genomics*, **11**(1), 107.

Arazi, H., Yoselewitz, I., Malka, Y., Kelner, Y., Genin, O. and Pines, M. (2009) Osteopontin and calbindin gene expression in the eggshell gland as related to eggshell abnormalities. *Poult Sci*, **88**, 647–653

Arora, R., Bhatia, S., Mishra, B. P., Jain, A. and Prakash, B. (2011) Diversity analysis of sheep breeds from Southern peninsular and Eastern regions of India, *Tropical Animal Health and Production*, **43**(2), 401-408.

Auer, H., Mayr, B., Lambrou, M. and Schleger, W. (1987) An extended chicken karyotype, including the NOR chromosome, *Cytogenetic and Genome Research*, **45**(3-4), 218-221.

Baker, J., Hardy, M. P., Zhou, J., Bondy, C., Lupu, F., Bellvé, A. R. and Efstratiadis, A. (1996) Effects of an Igf1 gene null mutation on mouse reproduction, *Molecular Endocrinology*, **10**(7), 903-918.

Barbato, G. (1999) Genetic relationships between selection for growth and reproductive effectiveness, *Poultry Science*, **78**(3), 444-452.

Baron, E. E., Moura, A. S. A. M. T., Ledur, M. C., Pinto, L. F. B., Boschiero, C., Ruy, D. C., Nones, K., Zanella, E. L., Rosário, M. F., Burt, D. W. and Coutinho, L. L. (2011) QTL for percentage of carcass and carcass parts in a broiler x layer cross, *Animal Genetics*, **42**(2), 117-124.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**(2), 263-265

Beaumont, M. A. and Balding, D. J. (2004) Identifying adaptive genetic divergence among populations from genome scans, *Molecular Ecology*, **13**(4), 969-980.

Beaumont, M. A. and Nichols, R. A. (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure, *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **263**(1377), 1619-1626.

Biswas, S. and Akey, J. M. (2006) Genomic insights into positive selection, *Trends in Genetics*, **22**(8), 437-446.

Bitgood, M. J., Shen, L. and McMahon, A. P. (1996) Sertoli cell signaling by Desert hedgehog regulates the male germline, *Current Biology*, **6**(3), 298-304.

Boitard, S. and Rocha, D. (2013) Detection of signatures of selective sweeps in the Blonde d'Aquitaine cattle breed, *Animal Genetics*, **44**(5), 579-583.

Boitard, S., Schlötterer, C. and Futschik, A. (2009) Detecting Selective Sweeps: A New Approach Based on Hidden Markov Models, *Genetics*, **181**(4), 1567-1578.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S. and SanCristobal, M. (2010) Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended, *Genetics*, **186**(1), 241-262.

Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K. and Cavalli-Sforza, L. L. (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms, *Proceedings of the National Academy of Sciences*, **88**(3), 839-843.

Brann, D. W. and Mahesh, V. B. (1997) Excitatory amino acids: evidence for a role in the control of reproduction and anterior pituitary hormone secretion. *Endocr Rev*, **18**, 678–700.

Burt, D. W. (2005) Chicken genome: Current status and future opportunities, *Genome Research*, **15**(12), 1692-1698.

Cabrera, C. P., Navarro, P., Huffman, J. E., Wright, A. F., Hayward, C., Campbell, H., Wilson, J. F., Rudan, I., Hastie, N. D., Vitart, V. and Haley, C. S. (2012) Uncovering Networks from Genome-Wide Association Studies via Circular Genomic Permutation, *G3: Genes|Genomes|Genetics*, **2**(9), 1067-1075.

Campos, R. L. R., Nones, K., Ledur, M. C., Moura, A. S. A. M. T., Pinto, L. F. B., Ambo, M., Boschiero, C., Ruy, D. C., Baron, E. E., Ninov, K., Altenhofen, C. A. B., Silva, R. A. M. S., Rosário, M. F., Burt, D. W. and Coutinho, L. L. (2009) Quantitative trait loci associated with fatness in a broiler–layer cross, *Animal Genetics*, **40**(5), 729-736.

Charlesworth, B. (1998) Measures of divergence between populations and the effect of forces that reduce variability, *Molecular Biology and Evolution,* **15(**5), 538-543.

Charlesworth, B. and Charlesworth, D. (2010) *Elements of Evolutionary Genetics.* Roberts and Company, p. 413.

Chadt, A., Leicht, K., Deshmukh, A., Jiang, L. Q., Scherneck, S., Bernhardt, U., Dreja, T., Vogel, H., Schmolz, K., Kluge, R., Zierath, J. R., Hultschig, C., Hoeben, R. C., Schürmann, A., Joost H. G. and Al-Hasani, H. (2008) Tbc1d1 mutation in lean mouse strain confers leanness and protects from diet-induced obesity. *Nature Genetics*, **40**, 1354–1359

Chen, H., Patterson, N. and Reich, D. (2010) Population differentiation as a test for selective sweeps, *Genome Research*, **20**(3), 393-402.

Choquet H., Stijnen P. and Creemers J. W. (2011) Genetic and Functional Characterization of PCSK1. *Methods Mol Biol* **768**, 247–253.

Clark, A. M., Garland, K. K. and Russell, L. D. (2000) Desert hedgehog (Dhh) Gene Is Required in the Mouse Testis for Formation of Adult-Type Leydig Cells and Normal Development of Peritubular Cells and Seminiferous Tubules, *Biology of Reproduction*, **63**(6), 1825-1838.

Crawford, R.D. (1990) Poultry Breeding and Genetics (Ed. Crawford, R.D.). Elsevier Science Publishing Company, Amsterdam and New York

Cruickshank, T. E. and Hahn, M. W. (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow, *Molecular Ecology*, **23**(13), 3133-3157.

Delany, M. E., Robinson, C. M., Goto, R. M. and Miller, M. M. (2009) Architecture and Organization of Chicken Microchromosome 16: Order of the NOR, MHC-Y, and MHC-B Subregions, *Journal of Heredity*, **100**(5), 507-514

Dunnington, E. A., Stallard, L. C., Hillel, J. and Siegel, P. B. (1994) Genetic Diversity Among Commercial Chicken Populations Estimated from DNA Fingerprints. *Poultry Science*, **73**, 1218-1225.

Elferink, M., van As, P., Veenendaal, T., Crooijmans, R. and Groenen, M. (2010) Regional differences in recombination hotspots between two chicken populations, *BMC Genetics*, **11**(1), 11.

Elferink, M. G., Megens, H.-J., Vereijken, A., Hu, X., Crooijmans, R. P. M. A. and Groenen, M. A. M. (2012) Signatures of Selection in the Genomes of Commercial and Non-Commercial Chicken Breeds, *PLoS ONE*, **7**(2), e32720.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'hom, B., Tixier-Boichard, M., Strömstedt, L., Wright, D., Jungerius, A., Vereijken, A., Randi, E., Jensen, P. and Andersson, L. (2008) Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. *PLoS Genet*, **4**(2), e1000010

Fan, W.-L., Ng, C. S., Chen, C.-F., Lu, M.-Y. J., Chen, Y.-H., Liu, C.-J., Wu, S.-M., Chen, C.-K., Chen, J.-J., Mao, C.-T., Lai, Y.-T., Lo, W.-S., Chang, W.-H. and Li, W.-H.

(2013) Genome-Wide Patterns of Genetic Variation in Two Domestic Chickens, *Genome Biology and Evolution*, **5**(7), 1376-1392.

Fay, J. C. and Wu, C.-I. (2000) Hitchhiking Under Positive Darwinian Selection, *Genetics*, **155**(3), 1405-1413.

Fillon, V. (1998) The chicken as a model to study microchromosomes in birds: a review, *Genetics Selection Evolution*, **30**, 209-212.

Flori, L., Fritz, S., Jaffrézic, F., Boussaha, M., Gut, I., Heath, S., Foulley, J.-L. and Gautier, M. (2009) The Genome Response to Artificial Selection: A Case Study in Dairy Cattle, *PLoS ONE*, **4**(8), e6595.

Fu, Y. X. and Li, W. H. (1993) Statistical tests of neutrality of mutations, *Genetics*, **133**(3), 693-709.

Fullerton, S. M., Bartoszewicz, A., Ybazeta, G., Horikawa, Y., Bell, G. I., Kidd, K. K., Cox, N. J., Hudson, R. R. and Di Rienzo, A. (2002) Geographic and Haplotype Structure of Candidate Type 2 Diabetes-Susceptibility Variants at the Calpain-10 Locus, *The American Journal of Human Genetics*, **70**(5), 1096-1106.

Fumihito, A., Miyake, T., Sumi, S., Takada, M., Ohno, S. and Kondo, N. (1994) One subspecies of the red junglefowl (Gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds, *Proceedings of the National Academy of Sciences*, **91**(26), 12505-12509.

Gerdina, A. K., Survea, V. V., Jönssona, M., Bjursella, M., Björkmana, M., Edenroa, A., Schuelkea, M., Saadb, A., Bjurströmb, S., Lundgrenb, E. J., Snaitha, M., Fransson-Steenb, R., Törnella, J., Bergb, A. L. and Bohlooly-Y, M. (2006) Phenotypic screening of hepatocyte nuclear factor (HNF) 4-γ receptor knockout mice. *Biochemical and Biophysical Research Communications*, **349**, 825–832.

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., Kappes, S. M., Lien, S., Matukumalli, L. K., McEwan, J. C., Nazareth, L. V., Schnabel, R. D., Weinstock, G. M., Wheeler, D. A., Ajmone-Marsan, P., Boettcher, P. J., Caetano, A. R., Garcia, J. F., Hanotte, O., Mariani, P., Skow, L. C., Sonstegard, T. S., Williams, J. L., Diallo, B., Hailemariam, L., Martinez, M. L., Morris, C. A., Silva, L. O., Spelman, R. J., Mulatu, W., Zhao, K., Abbey, C. A., Agaba, M., Araujo, F. R., Bunch, R. J., Burton, J., Gorni, C., Olivier, H., Harrison, B. E., Luff, B., Machado, M. A., Mwakaya, J., Plastow, G., Sim, W., Smith, T., Thomas, M. B., Valentini, A., Williams, P., Womack, J., Woolliams, J. A., Liu, Y., Qin, X., Worley, K. C., Gao, C., Jiang, H., Moore, S. S., Ren, Y., Song, X. Z., Bustamante, C. D., Hernandez, R. D., Muzny, D. M., Patil, S., San Lucas, A., Fu, Q., Kent, M. P., Vega, R., Matukumalli, A., McWilliam, S., Sclep, G., Bryc, K., Choi, J., Gao, H., Grefenstette, J. J., Murdoch, B., Stella, A., Villa-Angulo, R., Wright, M., Aerts, J., Jann, O., Negrini, R., Goddard, M. E., Hayes, B. J., Bradley, D. G., Barbosa da Silva, M., Lau, L. P., Liu, G. E., Lynn, D. J., Panzitta, F. and Dodds, K. G. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds, *Science (New York, N.Y.)*, **324**(5926), 528-532.

Gjederm, T. (1979) Selection for growth rate and domestication in Atlantic salmon, *Zeitschrift für Tierzüchtung und Züchtungsbiologie*, **96**(1-4), 56-59.

Godfrey, P., Rahal, J. O., Beamer, W. G., Copeland, N. G., Jenkins, N. A. and Mayo, K. E. (1993) GHRH receptor of little mice contains a missense mutation in the extracellular domain that disrupts receptor function, *Nature Genetics,* **4**(3), 227-232.

Granevitze, Z., Hillel, J., Feldman, M., Six, A., Eding, H. and Weigend, S. (2009) Genetic structure of a wide-spectrum chicken gene pool, *Animal Genetics,* **40**(5), 686-693.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Project, T. G. and Bustamante, C. D. (2011) Demographic history and rare allele sharing among human populations, *Proceedings of the National Academy of Sciences*.

Grobet, L., Martin, L. J. R., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., Schoeberlein, A., Dunner, S., Ménissier, F., Massabanda, J., Fries, R., Hanset, R. and Georges, M. (1997) A deletion in the bovine myostatin gene causes the double−muscled phenotype in cattle, *Nature Genetics,* **17**, 71-74.

Grossman, Sharon R., Andersen, Kristian G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, Daniel J., Griesemer, D., Karlsson, Elinor K., Wong, Sunny H., Cabili, M., Adegbola, Richard A., Bamezai, Rameshwar N. K., Hill, Adrian V. S., Vannberg, Fredrik O., Rinn, John L., Lander, Eric S., Schaffner, Stephen F. and Sabeti, Pardis C. (2013) Identifying Recent Adaptations in Large-Scale Genomic Data, *Cell,* **152**(4), 703-713.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F. and Sabeti, P. C. (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection, *Science*, **327**(5967), 883-886.

Gutiérrez-Gil, B., Arranz, J. J., Pong-Wong, R., García-Gámez, E., Kijas, J. and Wiener, P. (2014) Application of Selection Mapping to Identify Genomic Regions Associated with Dairy Production in Sheep, *PloS One*, **9**(5), e94623.

Haley, C. S. (1995) Livestock QTLs — bringing home the bacon?, *Trends in Genetics,* **11**(12), 488-492.

Hanon, E. A., Lincoln, G. A., Fustin, J. M., Dardente, H., Masson-Pévet, M., Morgan, P. J. and Hazlerigg, D. G. (2008) Ancestral TSH Mechanism Signals Summer in a Photoperiodic Mammal, *Curr. Biol*., **18**, 1147–1152

Harper, P. A. W., Latter, M. R., Nicholas, F. W., Cook, R. W. and Gill, P. A. (1998) Chondrodysplasia in Australian Dexter cattle, *Australian veterinary journal,* **76**(3), 199-202.

Hayes, B. J., Lien, S., Nilsen, H., Olsen, H. G., Berg, P., Maceachern, S., Potter, S. and Meuwissen, T. H. E. (2008) The origin of selection signatures on bovine chromosome 6, *Animal Genetics*, **39**(2), 105-111.

Herzog, W., Zeng, X., Lele, Z., Sonntag, C., Ting, J.-W., Chang, C.-Y. and Hammerschmidt, M. (2003) Adenohypophysis formation in the zebrafish and its dependence on sonic hedgehog, *Developmental Biology,* **254**(1), 36-49.

Hill, W. G. and Robertson, A. (1968) Linkage disequilibrium in finite populations, *Theoretical and Applied Genetics,* **38**(6), 226-231.

Hocking, P. M., Morrice, D. M., Law, A. S. and Burt, D. W. (2012) Many quantitative trait loci for feather growth in an F2 broiler × layer cross collocate with body weight loci. *British Poultry Science*, **53**(2), 162-167

Hu, Z.-L., Park, C. A., Wu, X.-L. and Reecy, J. M. (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era, *Nucleic Acids Research*, **41**(D1), D871-D879.

Huang, H. Y., Li, S.F., Zhao, Z.H., Liang, Z., Zhang, J. and Ding, Y. R. (2011) Association of polymorphisms for nuclear receptor coactivator 1 gene with egg production traits in the maternal line of Shaobo hens. *Br Poult Sci*, **52**, 328–332

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, *Nature*, **432**(7018), 695-716.

Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.-M., Beggs, W., Hoffman, G., Mezey, J. and Tishkoff, S. A. (2012) Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies, *PLoS Genet*, **8**(4), e1002641.

Jensen, J. D. (2014) On the unfounded enthusiasm for soft selective sweeps, *Nat Commun*, **5**.

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F. and Bustamante, C. D. (2005) Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data, *Genetics*, **170**(3), 1401-1410.

Jiao, S., Ren, H., Li, Y., Zhou, J., Duan, C. and Lu, L. (2013) Differential regulation of IGF-I and IGF-II gene expression in skeletal muscle cells. *Mol Cell Biochem*, **373**, 107–113

Johnson, J. A., Dunn, P. O. and Bouzat, J. L. (2007) Effects of recent population bottlenecks on reconstructing the demographic history of prairie-chickens, *Molecular Ecology*, **16**(11), 2203-2222.

Johnston, D. J., Herd, R., Reverter, A. and Oddy, V. H. (2001) Heritability of IGF-I in beef cattle and its association with growth and carcase traits, *Proc. Assoc. Adv. Anim. Breed. Genet*, **14**, 163-166.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K., Paiva, S., Barendse, W., Ciani, E., Raadsma, H., McEwan, J., Dalrymple, B. and other members of the International Sheep Genomics, C. (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection, *PLoS Biol*, **10**(2), e1001258.

Kim, E.-S., Cole, J. B., Huson, H., Wiggans, G. R., Van Tassell, C. P., Crooker, B. A., Liu, G., Da, Y. and Sonstegard, T. S. (2013) Effect of Artificial Selection on Runs of Homozygosity in U.S. Holstein Cattle, *PloS One*, **8**(11), e80813.

Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome., *Genetics*, **160**(2), 765–777.

Kindt, A. S. D., Navarro, P., Semple, C. A. M. and Haley, C. (2013) The genomic signature of trait-associated variants, *BMC Genomics*, **14**(1), 108.

Kranis, A., Gheyas, A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., Kaiser, P., Hocking, P., Fife, M., Salmon, N., Fulton, J., Strom, T., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., Qanbari, S., Simianer, H., Watson, K., Woolliams, J. and Burt, D. (2013) Development of a high density 600K SNP genotyping array for chicken, *BMC Genomics*, **14**(1), 59.

Lachance, J. and Tishkoff, S. A. (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it, *Bioessays*, **35**(9), 780-786.

Lewontin, R. C. (1964) THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS, *Genetics*, **49**(1), 49-67.

Lewontin, R. C. and Krakauer, J. (1973) DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS, *Genetics*, **74**(1), 175-195.

Li, D. F., Liu, W. B., Liu, J. F., Yi, G. Q., Lian, L., Qu, L. J., Li, J. Y., Xu, G. Y. and Yang, N. (2012) Whole-genome scan for signatures of recent selection reveals loci associated with important traits in White Leghorn chickens, *Poultry Science*, **91**(8), 1804-1812.

Li, D. Y., Zhang, L., Smith, D. G., Xu, H. L., Liu, Y. P., Zhao, X. L., Wang, Y. and Zhu, Q. (2013) Genetic effects of melatonin receptor genes on chicken reproductive traits, *Czech J. Anim. Sci.*, **58**(2), 58–64.

Litingtung, Y. and Chiang, C. (2000) Control of Shh activity and signaling in the neural tube, *Developmental Dynamics*, **219**(2), 143-154.

Liu, J.-P., Baker, J., Perkins, A. S., Robertson, E. J. and Efstratiadis, A. (1993) Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r), *Cell*, **75**(1), 59-72.

Ma, Y., Zhang, H., Zhang, Q. and Ding, X. (2014) Identification of Selection Footprints on the X Chromosome in Pig, *PLoS ONE*, **9**(4), e94911.

May, P., Woldt, E., Matz, R. L. and Boucher, P. (2007) The LDL receptor-related protein (LRP) family: an old family of proteins with new physiological functions. *Ann Med*, **39**(3):219-228.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P. and Cierco-Ayrolles, C. (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness, *Heredity*, **108**(3), 285-291.

Matzke, M. A., Varga, F., Berger, H., Schernthaner, J., Schweizer, D., Mayr, B. and Matzke, A. J. M. (1990) A 41–42 bp tandemly repeated sequence isolated from nuclear envelopes of chicken erythrocytes is located predominantly on microchromosomes, *Chromosoma*, **99**(2), 131-137.

Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene, *Genetical Research*, **23**(01), 23-35.

McDerment, N. A., Wilson, P. W., Waddington, D., Dunn, I. C. and Hocking, P. M. (2012) Identification of novel candidate genes for follicle selection in the broiler breeder ovary, *BMC Genomics*, **13**, 494

McVean, G. (2007) The Structure of Linkage Disequilibrium Around a Selective Sweep, *Genetics*, **175**(3), 1395-1406.

Moioli, B., Scatà, M. C., Steri, R., Napolitano, F. and Catillo, G. (2013) Signatures of selection identify loci associated with milk yield in sheep, *BMC Genetics*, **14**(76), doi:10.1186/1471-2156-14-76.

Moreng, R. E. and Avens, J. S. (1985) Poultry science and production. Reston Publishing Company, Inc. Reston, Virginia

Muir, W. M., Wong, G. K.-S., Zhang, Y., Wang, J., Groenen, M. A. M., Crooijmans, R. P. M. A., Megens, H.-J., Zhang, H., Okimoto, R., Vereijken, A., Jungerius, A., Albers, G. A. A., Lawley, C. T., Delany, M. E., MacEachern, S. and Cheng, H. H. (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds, *Proceedings of the National Academy of Sciences*, **105**(45), 17312-17317.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**(5881), 1344-9.

Need, A.C., Ahmadi, K.R., Spector, T.D. and Goldstein, D.B. (2006) Obesity is associated with genetic variants that alter dopamine availability. *Ann Hum Genet*, **70**, 293-303.

Nei, M. 1987. Molecular Evolutionary Genetics. New York: Columbia University Press.

Nei, M. and Maruyama, T. (1975) Lewontin-Krakauer test for neutral genes, *Genetics*, **80**(2), 395.

Nei, M., Maruyama, T. and Chakraborty, R. (1975) The Bottleneck Effect and Genetic Variability in Populations, *Evolution*, **29**(1), 1-10.

Ng, P. C. and Kirkness, E. F. (2010) Whole genome sequencing. *Methods Mol Biol*, **628**, 215-26

Nguyen, H., Danilenko, D., Bucay, N., DeRose, M., Van, G., Thomason, A. and Simonet, W. (1996) Expression of keratinocyte growth factor in embryonic liver of transgenic mice causes changes in epithelial growth and differentiation resulting in polycystic kidneys and other organ malformations., *Oncogene* **12**(10), 2109-2119.

Nielsen, J., Christiansen, J., Lykke-Andersen, J., Johnsen, A. H., Wewer, U. M. and Nielsen, F. C. (1999) A Family of Insulin-Like Growth Factor II mRNA-Binding Proteins Represses Translation in Late Development, *Molecular and Cellular Biology*, **19**(2), 1262-1270.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data, *Genome Research*, **15**(11), 1566-1575.

Nones, K., Ledur, M. C., Ruy, D. C., Baron, E. E., Melo, C. M. R., Moura, A. S. A. M. T., Zanella, E. L., Burt, D. W. and Coutinho, L. L. (2006) Mapping QTLs on chicken chromosome 1 for performance and carcass traits in a broiler x layer cross, *Animal Genetics*, **37**(2), 95-100.

Nones, K., Ledur, M. C., Zanella, E. L., Klein, C., Pinto, L. F. B., Moura, A. S. A. M. T., Ruy, D. C., Baron, E. E., Ambo, M., Campos, R. L. R., Boschiero, C., Burt, D. W. and Coutinho, L. L. (2012) Quantitative trait loci associated with chemical composition of the chicken carcass, *Animal Genetics*, **43**(5), 570-576.

Nusslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in Drosophila, *Nature*, **287**(5785), 795-801.

Oleksyk, T. K., Zhao, K., De La Vega, F. M., Gilbert, D. A., O'Brien, S. J. and Smith, M. W. (2008) Identifying Selected Regions from Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two Human Populations, *PLoS ONE*, **3**(3), e1712.

Parker, H. G., VonHoldt, B. M., Quignon, P., Margulies, E. H., Shao, S., Mosher, D. S., Spady, T. C., Elkahloun, A., Cargill, M., Jones, P. G., Maslen, C. L., Acland, G. M., Sutter, N. B., Kuroki, K., Bustamante, C. D., Wayne, R. K. and Ostrander, E. A. (2009) An Expressed Fgf4 Retrogene Is Associated with Breed-Defining Chondrodysplasia in Domestic Dogs, *Science*, **325**(5943), 995-998.

Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X., Tao, X., Wu, T., Ouzhuluobu, Basang, Ciwangsangbu, Danzengduojie, Chen, H., Shi, H. and Su, B. (2010) Genetic variations in Tibetan populations and high altitude adaptation at the Himalayas, *Molecular Biology and Evolution*.

Podisi, B. K., Knott, S. A., Dunn, I. C., Burt, D. W. and Hocking, P. M. (2012) Bone mineral density QTL at sexual maturity and end of lay, *British Poultry Science*, **53**(6), 763-769.

Podisi, B. K., Knott, S. A., Dunn, I. C., Law, A. S., Burt, D. W. and Hocking, P. M. (2011) Overlap of quantitative trait loci for early growth rate, and for body weight and age at onset of sexual maturity in chickens, *Reproduction*, **141**(3), 381-389.

Powell, J. E., Kranis, A., Floyd, J., Dekkers, J. C. M., Knott, S. and Haley, C. S. (2012) Optimal use of regression models in genome-wide association studies, *Animal Genetics*, **43**(2), 133-143.

Przeworski, M., Coop, G., Wall, J. D. and Nachman, M. (2005) The signature of positve selection on standing genetic variation, *Evolution*, **59**(11), 2312-2323.

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., Thaller, G. and Simianer, H. (2011) Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle, *BMC Genomics*, **12**(318), doi:10.1186/1471-2164-12-318.

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T. M., Fries, R., Nielsen, R. and Simianer, H. (2014) Classic Selective Sweeps Revealed by Massive Sequencing in Cattle, *PLoS Genet*, **10**(2), e1004148.

Qanbari, S., Strom, T. M., Haberer, G., Weigend, S., Gheyas, A. A., Turner, F., Burt, D. W., Preisinger, R., Gianola, D. and Simianer, H. (2012) A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens, *PLoS ONE*, **7**(11), e49525.

Randhawa, I. A., Khatkar, M., Thomson, P. and Raadsma, H. (2014) Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep, *BMC Genetics*, **15**(1), 34.

Resnyk, C., Carre, W., Wang, X., Porter, T., Simon, J., Le Bihan-Duval, E., Duclos, M., Aggrey, S. and Cogburn, L. (2013) Transcriptional analysis of abdominal fat in genetically fat and lean chickens reveals adipokines, lipogenic genes and a link between hemostasis and leanness, *BMC Genomics*, **14**(1), 557.

Riddle, R. D., Johnson, R. L., Laufer, E. and Tabin, C. (1993) Sonic hedgehog mediates the polarizing activity of the ZPA, *Cell*, **75**(7), 1401-1416.

Romanov, M. N. and Weigend S. (2001) Analysis of genetic relationships between various populations of domestic and jungle fowl using microsatellite markers. *Poult Sci*., **80**, 1057-63.

Rubin, C.-J., Megens, H.-J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C. B., Archibald, A. L., Fredholm, M., Groenen, M. A. M. and Andersson, L. (2012) Strong signatures of selection in the domestic pig genome, *Proceedings of the National Academy of Sciences*, **109**(48), 19529-19536.

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallbook, F., Besnier, F., Carlborg, O., Bed/'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K. and Andersson, L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication, *Nature*, **464**, 587-591.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E. S. (2002) Detecting recent positive selection in the human genome from haplotype structure, *Nature*, **419**(6909), 832 - 837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F. and Lander, E. S. (2007) Genome-wide detection and characterization of positive selection in human populations, *Nature*, **449**(7164), 913-918

Schütz, K., Kerje, S., Carlborg, O., Jacobsson, L., Andersson, L. and Jensen, P. (2002) QTL analysis of a red junglefowl x White Leghorn intercross reveals trade-off in resource allocation between behavior and production traits. *Behav Genet*, **32**, 423-33.

Schütz, K., Kerje, S., Jacobsson, L., Forkman, B., Carlborg, Ö., Andersson, L. and Jensen, P. (2004) Major Growth QTLs in Fowl Are Related to Fearful Behavior: Possible Genetic Links Between Fear Responses and Production Traits in a Red Junglefowl × White Leghorn Intercross, *Behavior Genetics*, **34**(1), 121-130.

Sewalem, A., Morrice, D., Law, A., Windsor, D., Haley, C., Ikeobi, C., Burt, D. and Hocking, P. (2002) Mapping of quantitative trait loci for body weight at three, six, and nine weeks of age in a broiler layer cross, *Poultry Science*, **81**(12), 1775-1781.

Sheffield, V. C. (2004) Use of isolated populations in the study of a human obesity syndrome, the Bardet-Biedl syndrome. *Pediatr Res*, **55**(6), 908-911.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Research*, **29**(1), 308-311.

Shia, Y., Suhb, Y. H., Milsteina, A. D., Isozakib, K., Schmida, S. M., Rocheb, K. W. and Nicolla, R .A. (2010) Functional comparison of the effects of TARPs and cornichons on AMPA receptor trafficking and gating. *PNAS*, **107**(37), 16315 - 16319

Shiratori, M., Oshika, E., Ung, L. P., Singh, G., Shinozuka, H., Warburton, D., Michalopoulos, G. and Katyal, S. L. (1996) Keratinocyte growth factor and embryonic rat lung morphogenesis, *American Journal of Respiratory Cell and Molecular Biology*, **15**(3), 328-338.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M. and Jones, K. W. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs, *Human Genomics*, **1**, 274-286.

Siegel, P. B., Haberfeld, A., Mukherjee, T. K., Stallard, L. C., Marks, H. L., Anthony, N. B. and Dunnington, E. A. (1992) Jungle fowl–domestic fowl relationships: a use of DNA fingerprinting. *World's Poultry Science Journal*, **48**, 147-155

Simianer, H. 'Statistical Problems in Livestock Population Genomics'. *10th World Congress on Genetics Applied to Livestock Production*: Asas.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T. and Ge, R. (2010) Genetic Evidence for High-Altitude Adaptation in Tibet, *Science*, **329**(5987), 72-75.

Smith, J., Bruley, C. K., Paton, I. R., Dunn, I., Jones, C. T., Windsor, D., Morrice, D. R., Law, A. S., Masabanda, J., Sazanov, A., Waddington, D., Fries, R. and Burt, D. W. (2000) Differences in gene density on chicken macrochromosomes and microchromosomes, *Animal Genetics*, **31**(2), 96-103.

Smith, J. and Burt, D. W. (1998) Parameters of the chicken genome (Gallus gallus), *Animal Genetics*, **29**(4), 290-294.

Solinhac, R., Leroux, S., Galkina, S., Chazara, O., Feve, K., Vignoles, F., Morisson, M., Derjusheva, S., Bed'hom, B., Vignal, A., Fillon, V. and Pitel, F. (2010) Integrative mapping analysis of chicken microchromosome 16 organization, *BMC Genomics* **11**, 616.

Stella, A., Ajmone-Marsan, P., Lazzari, B. and Boettcher, P. (2010) Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production, *Genetics*, **185**(4), 1451-1461.

Sun, Y., Zhao, G., Liu, R., Zheng, M., Hu, Y., Wu, D., Zhang, L., Li, P. and Wen, J. (2013) The identification of 14 new genes for meat quality traits in chicken using a genome-wide association study, *BMC Genomics*, **14**(1), 458.

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P. G., Quignon, P., Johnson, G. S., Parker, H. G., Fretwell, N., Mosher, D. S., Lawler, D. F., Satyaraj, E., Nordborg, M., Lark, K. G., Wayne, R. K. and Ostrander, E. A. (2007) A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs, *Science*, **316**(5821), 112-115.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics*, **123**(3), 585-95.

Tang, K., Thornton, K. R. and Stoneking, M. (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome, *PLoS Biology*, **5**(7), e171.

Tang, S., Sun, D., Ou, J., Zhang, Y., Xu, G. and Zhang, Y. (2010) Evaluation of the IGFs (IGF1 and IGF2) genes as candidates for growth, body measurement, carcass, and reproduction traits in Beijing You and Silkie chickens. *Anim Biotechnol*, **21**, 104–113

Taylor, A. C., Horvat-Gordon, M., Moore, A. and Bartell, P. A. (2013) The Effects of Melatonin on the Physical Properties of Bones and Egg Shells in the Laying Hen, *PloS One*, **8**(2), e55663.

Templeton, A. R. (1980) THE THEORY OF SPECIATION VIA THE FOUNDER PRINCIPLE, *Genetics*, **94**(4), 1011-1038.

Thesleff, I. (2003) Epithelial-mesenchymal signalling regulating tooth morphogenesis, *Journal of Cell Science*, **116**(9), 1647-1648.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A. and Deloukas, P. (2007) Convergent adaptation of human lactase persistence in Africa and Europe, *Nat Genet*, **39**(1), 31-40.

Tishkoff, S. A. and Verrelli, B. C. (2003) PATTERNS OF HUMAN GENETIC DIVERSITY: Implications for Human Evolutionary History and Disease, *Annual Review of Genomics and Human Genetics*, **4**(1), 293-340.

Twito, T., Madeleine, D., Perl-Treves, R., Hillel, J. and Lavi, U. (2011) Comparative genome analysis with the human genome reveals chicken genes associated with fatness and body weight, *Animal Genetics*, **42**(6), 642-649.

Utsunomiya, Y. T., Pérez O'Brien, A. M., Sonstegard, T. S., Van Tassell, C. P., do Carmo, A. S., Mészáros, G., Sölkner, J. and Garcia, J. F. (2013) Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods, *PloS One*, **8**(5), e64280.

Väisänen, J. and Jensen, P. (2004) Responses of young red jungle fowl (Gallus gallus) and White Leghorn layers to familiar and unfamiliar social stimuli. *Poult Sci*, **83**, 335-43.

Väisänen, J., Håkansson, J. and Jensen, P. (2005) Social interactions in Red Junglefowl (Gallus gallus) and White Leghorn layers in stable groups and after re-grouping. *Br Poult Sci*, **46**, 156-68.

Voight, B. F., Kudaravalli, S., Wen, X. and Pritchard, J. K. (2006) A Map of Recent Positive Selection in the Human Genome, *PLoS Biol*, **4**(3), e72.

Vortkamp, A., Lee, K., Lanske, B., Segre, G. V., Kronenberg, H. M. and Tabin, C. J. (1996) Regulation of Rate of Cartilage Differentiation by Indian Hedgehog and PTH-Related Protein, *Science*, **273**(5275), 613-622.

Wang, E. T., Kodama, G., Baldi, P. and Moyzis, R. K. (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens, *Proceedings of the National Academy of Sciences of the United States of America*, **103**(1), 135-140.

Watterson, G. A. (1978) The homozygosity test of neutrality, *Genetics*, **88**(2), 405-417.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. and Hill, W. G. (2005) Measures of human population structure show heterogeneity among genomic regions, *Genome Research*, **15**(11), 1468-1476.

Weir, B. S. and Cockerham, C. C. (1984) Estimating F-Statistics for the Analysis of Population Structure, *Evolution*, **38**(6), 1358-1370.

West, B. and Zhou, B.-X. (1989) Did chickens go north? New evidence for domestication, *World's Poultry Science Journal*, **45**(03), 205-218.

Whitlock, M. C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach, *Journal of Evolutionary Biology*, **18**(5), 1368-1373.

Wiener, P. and Pong-Wong, R. (2011) A Regression-Based Approach to Selection Mapping, *Journal of Heredity*, **102**(3), 294-305.

Wilkinson, S., Lu, Z. H., Megens, H.-J., Archibald, A. L., Haley, C., Jackson, I. J., Groenen, M. A. M., Crooijmans, R. P. M. A., Ogden, R. and Wiener, P. (2013) Signatures of Diversifying Selection in European Pig Breeds, *PLoS Genet*, **9**(4), e1003453.

Wong, G., Liu, B., Wang, J., Zhang, Y., Yang, X., Zhang, Z., Meng, Q., Zhou, J., Li, D., Zhang, J., Ni, P., Li, S., Ran, L., Li, H., Zhang, J., Li, R., Li, S. and Zheng, H. (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms, *Nature*, **432**(7018), 717-722.

Wood, B. J., Archer, J. A. and van der Werf, J. H. J. (2004) Response to selection in beef cattle using IGF-1 as a selection criterion for residual feed intake under different Australian breeding objectives, *Livestock Production Science*, **91**(1–2), 69-81.

Wright, S. (1951) The genetical structure of populations, *Ann. Eugenics*, **15**(323-354).

Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis, *American Journal of Human Genetics*, **88**(1), 76-82.

Yang, Y., Drossopoulou, G., Chuang, P. T., Duprez, D., Marti, E., Bumcrot, D., Vargesson, N., Clarke, J., Niswander, L., McMahon, A. and Tickle, C. (1997) Relationship between dose, distance and time in Sonic Hedgehog-mediated regulation of anteroposterior polarity in the chick limb, *Development*, **124**(21), 4393-4404.

Zeder, M. A. (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact, *Proceedings of the National Academy of Sciences*.

Zeng, K., Fu, Y.-X., Shi, S. and Wu, C.-I. (2006) Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants, *Genetics*, **174**(3), 1431-1439.

Zeng, K., Shi, S. and Wu, C.-I. (2007) Compound Tests for the Detection of Hitchhiking Under Positive Selection, *Molecular Biology and Evolution*, **24**(8), 1898-1908.

Zhang, Z. R., Liu, Y. P., Jiang, X., Du, H. R. and Zhu, Q. (2008) Study on association of single nucleotide polymorphism of CAPN1 gene with muscle fibre and carcass traits in quality chicken populations, *J Anim Breed Genet*, **125**, 258–264.

Zhang, H., Hu, X., Wang, Z., Zhang, Y., Wang, S., Wang, N., Ma, L., Leng, L., Wang, S., Wang, Q., Wang, Y., Tang, Z., Li, N., Da, Y. and Li, H. (2012a) Selection Signature Analysis Implicates the *PC1/PCSK1* Region for Chicken Abdominal Fat Content, *PLoS ONE*, **7**(7), e40736.

Zhang, H., Wang, S.-Z., Wang, Z.-P., Da, Y., Wang, N., Hu, X.-X., Zhang, Y.-D., Wang, Y.-X., Leng, L., Tang, Z.-Q. and Li, H. (2012b) A genome-wide scan of selective sweeps in two broiler chicken lines divergently selected for abdominal fat content, *BMC Genomics*, **13**(1), 704.

Zhou, H., Mitchell, A. D., McMurtry, J. P., Ashwell, C. M. and Lamont, S. J. (2005) Insulin-like growth factor-I gene polymorphism associations with growth, body composition, skeleton integrity, and metabolic traits in chickens, *Poultry Science*, **84**(2), 212-219.

Zhu, L. and Bustamante, C. D. (2005) A Composite-Likelihood Approach for Detecting Directional Selection From DNA Sequence Data, *Genetics*, **170**(3), 1411-1421.